CONCEPTUAL KNOWLEDGE OF CONFIDENCE INTERVALS IN PSYCHOLOGY UNDERGRADUATE AND GRADUATE STUDENTS

NOELLE M. CROOKS
Broward College
ncrooks@broward.edu

ANNA N. BARTEL
University of Wisconsin - Madison
anbartel@wisc.edu

MARTHA W. ALIBALI
University of Wisconsin - Madison
martha.alibali@wisc.edu

ABSTRACT

In recent years, there have been calls for researchers to report and interpret confidence intervals (CIs) rather than relying solely on p-values. Such reforms, however, may be hindered by a general lack of understanding of CIs and how to interpret them. In this study, we assessed conceptual knowledge of CIs in undergraduate and graduate psychology students. CIs were difficult and prone to misconceptions for both groups. Connecting CIs to estimation and sample mean concepts was associated with greater conceptual knowledge of CIs. Connecting CIs to null hypothesis significance testing, however, was not associated with conceptual knowledge of CIs. It may therefore be beneficial to focus on estimation and sample mean concepts in instruction about CIs.

Keywords: *Statistics education research; Estimation; Conceptual understanding*

1. INTRODUCTION

1.1. CONCEPTUAL KNOWLEDGE OF STATISTICS

Reforms in statistics education have highlighted the importance of improving statistical reasoning—defined as “the way people reason with statistical ideas and make sense of statistical information” (e.g., Garfield & Chance, 2000, p. 101). One component of statistical reasoning is understanding important concepts (e.g., Garfield, 2003). In fact, some have asserted that improving statistical reasoning will require a shift towards more conceptual learning and away from rote memorization and computation (e.g., Moore, 1997). Here, we use the term *conceptual knowledge* to refer to understanding of general principles and relationships, one of the forms of conceptual knowledge identified in a recent review (Crooks & Alibali, 2014). Conceptual knowledge of statistics is thought to include an understanding of the *why* of statistics in addition to the *how*.

Many studies have argued for the importance of conceptual knowledge in statistics. Conceptual knowledge is thought to allow students to think more flexibly (e.g., Jones, Jones, & Vermette, 2011), transfer knowledge to novel problems (e.g., Bude, Imbos, van de Wiel, & Berger, 2011; Bude, van de Wiel, Imbos, & Berger, 2010; Paas, 1992), decide what type of analysis to use (e.g., Bude et al., 2010; Bude et al., 2011; Graham & Thomas, 2005), represent information accurately (e.g., Graham & Thomas, 2005; Hong & O’Neil, 1992), make accurate comparative judgments (Bisson, Gilmore, Inglis, & Jones, 2016), understand data (e.g., Garfield & Chance, 2000; Jones et al., 2011), interpret results (e.g., Gal & Garfield, 1997; Jones et al., 2011), and think critically (e.g., Garfield & Chance, 2000). Unfortunately, however, many traditional statistics classes do not typically promote a high level of conceptual knowledge (delMas, Garfield, Ooms, & Chance, 2007; Meletiou-Mavrotheris & Lee, 2002;

Pfannkuch, Wild, & Parsonage, 2012), and, more generally, learning concepts tends to be more challenging than learning procedures (Leppink, Broers, Imbos, van der Vleuten, & Berger, 2012).

In our view, learning statistics involves building on existing knowledge structures, as well as acquiring new knowledge via instruction or experience. Learners sometimes adapt their existing knowledge structures in ways that are inaccurate, but that may be functional, at least in some contexts (see, e.g., Smith, diSessa, & Roschelle, 1994). From this perspective, it is critical to understand both whether students have accurate conceptual knowledge and also whether they hold flawed, inaccurate, or incomplete conceptions, which we will refer to using the umbrella term, *misconceptions*.

In light of the value of conceptual knowledge in statistics, there is a need for research focusing on statistical topics for which conceptual difficulties are particularly widespread. Specifically, there is a need for research illuminating what exactly students know about conceptually difficult statistical topics and how statistics lessons can be structured to foster conceptual knowledge of such topics.

1.2. CONFIDENCE INTERVALS

In recent years, researchers have been encouraged to decrease their reliance on significance testing and *p*-values, and to focus more on estimation and practical significance (e.g., Cumming, Fidler, Kalinowski, & Lai, 2012; Cumming & Fidler, 2009). Indeed, the American Statistical Association recently released a statement acknowledging the misuse of *p*-values, discussing principles for their appropriate use and interpretation, and acknowledging alternative approaches that emphasize estimation rather than testing (ASA, 2016). One step in shifting away from significance testing is reporting and interpreting confidence intervals (CIs) in empirical work. In the field of psychology, the American Psychological Association (APA) has supported these efforts. In fact, the most recent (sixth) edition of the APA publication manual notes that, “because confidence intervals combine information on location and precision and can often be used directly to infer significance levels, they are, in general, the best reporting strategy” (2010, p. 34). Some journals now recommend that authors report CIs; for example, the author guidelines for *Psychological Science* (the flagship journal of the Association for Psychological Science) now “recommend the use of the ‘new statistics’—effect sizes, confidence intervals, and meta-analysis—to avoid problems associated with null-hypothesis significance testing” (APS, 2017).

Despite efforts by the APA and by journal editors, there has been little success getting psychological researchers to include CIs in their work, and even less success getting them to interpret CIs correctly (Cumming, 2014; Cumming et al., 2007; Fidler et al., 2005; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Even in the face of explicit efforts to implement statistical reforms, there has been little change in researchers’ practices (Cumming et al., 2007). More specifically, the statistical recommendations of the sixth edition of the APA’s publication manual have had a limited impact on practice (Cumming et al., 2012). One hypothesis about the resistance to CIs is that many researchers do not have a good grasp of their conceptual basis.

1.3. CONCEPTUAL KNOWLEDGE OF CONFIDENCE INTERVALS

Key concepts Conceptual knowledge of CIs involves the ability to accurately interpret the calculated interval and to relate it to other statistical concepts. Although there are many statistical concepts that relate to CIs in some way, a few specific concepts seem central to deep conceptual knowledge of CIs. A review of previous research, statistics textbooks, and video data from statistics instruction (Lockwood, Yeo, Crooks, Nathan, & Alibali, 2014) highlights some key concepts that appear to be particularly important when thinking about CIs. These concepts include (a) understanding the definition of the term “confidence interval,” (b) understanding the distinction between sample and population means and how they are related, (c) understanding the notion of confidence level (i.e., 90% vs. 95% CI), (d) understanding how various factors (e.g., sample size, sample variability) affect CI width, (e) understanding what can be inferred about future replications based on CIs, and (f) understanding how to interpret CIs accurately.

Unfortunately, past work suggests that CI-related concepts are difficult for researchers (e.g., Coulson, Healey, Fidler, & Cumming, 2010; Cumming, 2006). Some evidence suggests that CIs are difficult for students as well (Henriques, 2016). In developing a broad assessment of conceptual knowledge of statistics, delMas and colleagues (2007) noted that performance by undergraduate students on CI items was poor. Given the breadth of their assessment, there were only a small number of items that specifically focused on CIs, but the data did reveal participants' difficulty with CI concepts. Specifically, whereas 75% of the undergraduate sample was able to identify a correct interpretation of a confidence interval, many of these students endorsed an incorrect interpretation, as well. This suggests that these students believed that the correct and incorrect interpretations conveyed similar ideas, which is inaccurate. Furthermore, the majority of students in the sample demonstrated misconceptions about CIs, even after taking an introductory statistics course. Additionally, there is anecdotal evidence from instructors noting that CIs are a particularly difficult topic for students (Holte, 2003). Taken together, data from both advanced researchers and beginning statistics students suggest that understanding the conceptual basis of CIs presents a challenge.

Common misconceptions Understanding CIs involves not only knowledge of concepts, but also the absence of misconceptions. In this paper we use the term "misconceptions" to refer to incomplete or flawed conceptions that may nevertheless be functional or useful in some contexts (see Smith et al., 1994, for discussion). Several distinct misconceptions have been identified in previous research (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; Cumming & Maillardet, 2006; Fidler, 2006; Grant & Nathan, 2008; Greenland et al., 2016; Henriques, 2016); we focus here on five key misconceptions (Table 1). Specifically, the *Sample Mean Misconception* is the belief that a CI allows one to estimate the sample mean. The *Confidence Level Misconception* is the belief that the confidence level of a calculated interval indicates the percentage of replication means that will fall within the original interval. The *Individual Scores Misconception* is the belief that a CI gives the range of individual scores. The *Fixed Interval Misconception* is the belief that a CI is a fixed interval, within which a moving parameter may or may not fall. Finally, the *Equality Misconception* is the belief that a CI gives the likelihood of the sample mean being equal to the population mean.

Table 1. CI misconceptions

Misconception	Source	Explanation
Sample Mean (SM)	Castro Sotos et al. (2007); Fidler (2006)	A CI allows one to estimate the *sample* mean
Confidence Level (CL)	Castro Sotos et al. (2007); Cumming & Maillardet (2006)	The confidence level of a calculated interval indicates the percentage of replication means that will fall within the original interval; e.g., a 95% interval calculated from a set of data will include 95% of the sample means from all replications
Individual Scores (IS)	Castro Sotos et al. (2007); Fidler (2006)	A CI gives the range of individual scores; a CI gives the range of individual scores within some margin of error (e.g., 1 SD)
Fixed Interval (FI)	Grant & Nathan (2008)	A CI is a fixed interval, within which a moving parameter may or may not fall
Equality (EQ)	Fidler (2006)	A CI gives the likelihood of the sample mean being equal to the population mean

Past research has documented misconceptions about CIs in a range of populations, including researchers (e.g., Cumming, Williams, & Fidler, 2004; Hoekstra, Rouder, Morey & Wagenmakers, 2014), graduate students (e.g., Grant & Nathan, 2008; Hoekstra et al., 2014), and undergraduates (e.g., Fidler, 2006; Henriques, 2016; Reaburn, 2014). For example, in a study of researchers in a number of fields (including psychology), Cumming and colleagues (2004) asked participants to predict what would happen if the study were replicated. Performance was poor and riddled with misconceptions, even for participants who were actively publishing in fields in which reporting CIs is the norm (e.g., medicine). Researchers commonly underestimated the degree to which a parameter in a replication would vary from its originally reported value. In another study that included researchers, master's students, and undergraduates in psychology, Hoekstra and colleagues (2014) asked participants to indicate whether each of a set of false statements about CIs was true or false. In each of the three participant groups, participants endorsed more than half of the false statements, on average.

Interpreting confidence intervals CIs allow for two types of statistical inference: specifically, they can be used for both estimation and hypothesis testing. Deep conceptual knowledge of CIs involves the ability to use CIs for both estimation and hypothesis testing, as it is precisely this potential for inference that makes CIs so useful (e.g., APA, 2010). Although confidence intervals allow for testing any hypothesis about the relevant population parameter, we focus here on null hypothesis significance testing (NHST), given the prevalence of this type of hypothesis testing in the teaching and learning of statistics.

It has been suggested that thinking about CIs primarily in terms of NHST might be detrimental to conceptual knowledge of CIs (e.g., Cumming, 2012, 2014; Grant & Nathan, 2008). For example, in one study, researchers were asked to interpret data from two fake studies, one with significant findings and the other with non-significant findings, presented in a variety of formats (i.e., CIs, *p*-values). To correctly interpret the data, participants needed to recognize that the findings across the two studies were not inconsistent, despite the difference in statistical significance. The data showed that interpretations varied widely across participants and were not always correct (Coulson et al., 2010). Additionally, many participants interpreted data presented in CI format in terms of NHST, and those participants who gave NHST-based interpretations (regardless of the initial presentation format or the correctness of the NHST-based interpretations themselves) performed worse overall than those who interpreted the data without relying on such concepts. Thus, it appears that thinking about CIs primarily in terms of NHST might be associated with lower levels of conceptual knowledge, at least for researchers. Little research has specifically examined the relationship between NHST-based thinking and conceptual knowledge of CIs in students (see Fidler, 2005, for an exception). Additionally, no studies have specifically addressed the relationship between thinking about CIs primarily in terms of estimation and conceptual knowledge of CIs.

One key concept underlying CI interpretation is the distinction between samples and populations. As a tool for inferential statistics, CIs allow one to use sample data to make generalizations about the population. These types of generalizations, however, are contingent on a more general understanding of samples. No previous studies have examined how knowledge of the role of samples in making inferences about a population relates to conceptual knowledge of CIs. It seems likely that lack of understanding of sample mean concepts may be problematic for students' understanding of CIs (Fidler, 2006).

1.4. CURRENT STUDY

In brief, the goal of the current study was to assess undergraduate and graduate psychology students' conceptual knowledge of CIs. As discussed above, some past research has investigated knowledge of CIs in different samples of participants (e.g., delMas et al., 2007; Fidler, 2006; Fidler & Loftus, 2009; García-Pérez & Alcalá-Quintana, 2016; Grant & Nathan, 2008; Henriques, 2016; Hoekstra, Johnson, & Kiers, 2012; Hoekstra et al, 2014). Most prior studies, however, used only a few items to assess CI misconceptions or CI knowledge, and none included a comprehensive measure of conceptual knowledge of CIs. In order to assess participants' conceptual knowledge, we developed a set of questions designed to tap knowledge of CIs and key related concepts. Based on previous work, we expected that performance would be low and that misconceptions would be prevalent.

In addition to measuring students' conceptual knowledge, we were interested in how the content of students' answers, specifically their tendency to talk about CIs primarily in terms of either NHST or estimation, related to their conceptual knowledge of CIs. We predicted that the tendency to mention NHST in responses would be related to lower levels of conceptual knowledge (e.g., Coulson et al., 2010). Conversely, we predicted that the tendency to mention estimation in responses would be related to higher levels of conceptual knowledge. Finally, we predicted that strong understanding of the differences between sample and population means would be positively related to conceptual knowledge.

2. METHODS

2.1. PARTICIPANTS

The undergraduate sample included 21 students, all of whom had completed or were taking a Basic Statistics for Psychology course at a large university in the Midwestern United States. All participants who were enrolled in the course at the time of participation had already covered CIs in class. Participants ranged in age from 18–23 (*M* = 20 years, 6 months, *SD* = 1.47), and 80% were female. There were four freshman, six sophomores, seven juniors, three seniors, and one fifth-year student in the sample. Eighty percent of participants identified as White, and the remaining 20% identified as Asian. For most participants, Basic Statistics for Psychology was the only college-level statistics course they had taken; only four students had completed an additional statistics class. Overall, participants had high standardized test scores, ranging from the 74th–99th percentile on the quantitative portion of the SAT or ACT (*M* = 93.5, *SD* = 6.2).

The graduate sample consisted of 19 psychology graduate students, all of whom had completed at least one graduate-level statistics course. Participants represented a variety of areas of psychology (developmental, biological, clinical, cognitive, cognitive neuroscience, perceptual, and social). Participants ranged in age from 22–40 (*M* = 28 years, 1 month, *SD* = 4.2), and 58% were female. They had been in graduate school for one to seven years (*M* = 4 years; *SD* = 1.6). Seventy-four percent of the students identified as White, 16% as Asian, 5% as Black, and 5% as some other race or ethnicity. Although completion of only one graduate statistics course was required for participation, all participants had taken or were in the process of taking at least one additional course (*M* = 3.3, *SD* = 1.1). Additionally, all but one participant reported having taken at least one statistics class as an undergraduate. Nine participants had served as TAs for at least one undergraduate statistics or methods course in the psychology department during their time in graduate school. One participant had also served as a TA for the introductory graduate statistics course in the psychology department. In terms of standardized test scores, participants ranged from the 87th–99th percentile on the quantitative section of the ACT/SAT (*M* = 95.2, *SD* = 4.5) and from the 45th–93rd percentile on the quantitative section of the GRE (*M* = 74.3, *SD* = 16.8). It should be noted that these scores reflect high GRE performance. All participants had completed the GRE before its recent re-norming. Therefore, although six participants received a perfect GRE score, they were only in the 93rd percentile.

2.2. MATERIALS

Conceptual knowledge assessment instrument As a first step in creating the assessment instrument, we compiled a set of potential test items from three sources: Basic Statistics for Psychology instructors, undergraduate statistics textbooks, and previous research. We reviewed these items in light of the major concepts that underlie CIs. As stated above, these concepts were (a) understanding the definition of the term “confidence interval,” (b) understanding the distinction between sample and population means and how they are related, (c) understanding the notion of confidence level (i.e., 90% vs. 95% CI), (d) understanding how various factors (e.g., sample size, sample variability) affect CI width, (e) understanding what can be inferred about future replications based on CIs, and (f) understanding how to interpret CIs accurately. Fourteen items, including open-response, true/false, and multiple-choice questions, were then selected for piloting.

The initial fourteen-item test was piloted on seven professors/post-docs, seven graduate students, and ten undergraduates. Pilot participants were asked to provide responses to the questions and also to note any thoughts they had about the items, such as possible alternative interpretations. Based on the performance and feedback of these pilot participants, the initial set of items was revised. Some items were removed, others were reworded, and a few new items were created.

The final conceptual knowledge assessment instrument consisted of 12 items, all in either open-response, forced choice, or true/false format. One item asked participants to recall the formula for calculating CIs; this item was excluded from analysis (see below). Each of the remaining items was intended to assess at least one of the identified CI concepts or misconceptions. Nine of the items were used to create a measure of *conceptual knowledge of CIs* (Table 2). One additional item was used as part of a measure of *understanding of sample mean concepts*. Participants’ responses to the open-response and explanation items were also coded for mentions of *estimation* and mentions of *null hypothesis significance testing*, as well as mentions of specific misconceptions. These measures are described in greater detail below. Appendix A lists the items in the order in which they were presented to participants.

Table 2. Items from the Conceptual Knowledge Assessment used to assess conceptual knowledge of CIs

Item	Target concept	Source	Content coding
Open Response Items			
Define the term "confidence interval."	Definition	Kirk, 1999	NHST, EST, M
In a study of the effects of marijuana use during pregnancy, measurements on babies of mothers who used marijuana during pregnancy were compared to measurements on babies of mothers who did not. A 95% confidence interval for the difference in mean head circumference (non-use minus use) was 0.61 to 1.19 cm. What can be said from this statement about the hypothesis that the mean difference is zero?	NHST interpretation of CI	Ramsey & Schafer, 2002	EST, M
Imagine you are describing confidence intervals to a beginning statistics student. Explain how to interpret the following confidence interval: 95% confidence interval $8.5 < \mu <11.5$.	Interpretation of CI	Grant & Nathan, 2008	NHST, EST, M
True or False + Explanation Items			
True or False: If all other factors are held constant, an 80% confidence interval is wider than a 90% confidence interval. Please explain your choice.	Confidence level	Gravetter & Wallnau, 2013	NHST, EST, M
True or False: If all other factors are held constant, a confidence interval computed from a sample of $n = 25$ is wider than a confidence interval computed from a sample of $n = 100$. Please explain your choice.	Relation to sample size	Gravetter & Wallnau, 2013	NHST, EST, M
True or False: If all other factors are held constant, a confidence interval computed from a sample with high variability is narrower than a confidence interval computed from a sample with low variability. Please explain your choice.	Relation to sample variability	Not applicable	NHST, EST, M
Forced Choice Items: Explain how each of the following affects the width of a confidence interval			
Increasing the sample size	Relation to sample size	Gravetter & Wallnau, 2000	Not applicable
Increasing the sample variability	Relation to sample variability	Gravetter & Wallnau, 2000	Not applicable
Increasing the level of confidence (the percent confidence)	Confidence level	Gravetter & Wallnau, 2000	Not applicable

Note. NHST = null hypothesis significance testing, EST = estimation, M = misconceptions

Misconception assessment instrument A second instrument was created to gauge the existence of CI misconceptions, including the Sample Mean, Confidence Level, Individual Scores, Fixed Interval, and Equality misconceptions (Table 1). This instrument consisted of eight statements (Table 3) that participants were asked to rate on a 5-point scale ranging from *Very Inaccurate* to *Very Accurate*. Six of the statements were incorrect (embodying one or more CI misconceptions); two reflected accurate conceptions of CIs. Appendix B lists the items in the original order given to participants.

Table 3. Items on the Misconception Assessment

Concept or Misconception	Statement	Source
	Correct statements	
Definition	A 95% confidence interval is the interval for which you are 95% certain that it contains the population mean.	Not applicable
Replication	If you repeatedly take a sample of size *n* from a population and construct a 95% confidence interval each time, 95% of those intervals should contain the population mean.	Not applicable
	Misconception statements	
Sample Mean	A confidence interval gives you the range of possible values for the sample mean.	Fidler, 2006; Castro Sotos et al., 2007
Confidence Level, Sample Mean	If you were to conduct an infinite number of experiments exactly like the original experiment, a 95% confidence interval would contain 95% of the sample means from these experiments.	Cumming & Maillardet, 2006
Individual scores	A confidence interval gives you the range of the individual scores.	Fidler, 2006; Castro Sotos et al., 2007
Individual scores	A confidence interval gives you the range of the individual scores within one standard deviation of the population mean.	Fidler, 2006; Castro Sotos et al., 2007
Fixed Interval	A 95% confidence interval is the interval for which you are 95% confident that the population mean falls within it.	Based on Grant & Nathan, 2008
Equality	A 95% confidence interval indicates that there is a 95% chance that the sample mean equals the population mean.	Not applicable

Recall of the CI formula We also asked participants to recall the formula for calculating one-sample mean CIs and to label each part of the formula. Very few participants accurately recalled the formula (only 14% of undergraduates and 32% of graduate students). With this item, we had hoped to evaluate whether students' labels reflected accurate interpretations of the elements of the formula. We could not use this item to assess understanding in this way, however, because so few students recalled the formula, so we did not analyze this item further.

2.3. PROCEDURE

Students participated individually in a lab setting and were given up to an hour to complete the assessments. All participants completed the conceptual knowledge assessment first and the misconception assessment second. Approximately half of the participants (n = 11 undergraduates; n = 9 graduate students) completed the assessments in paper-and-pencil format and the remainder (n = 10 undergraduates; n = 10 graduate students) did so while talking aloud. Participants who were asked to talk aloud were videotaped. Assignment to test format was not entirely random, because not all participants consented to be recorded. Talk-aloud protocols were collected because we were interested in learning as much about participants' thought processes as possible, and participants often provide more information when speaking than when giving written responses (see Ericsson & Simon, 1993). Additionally, it has been suggested that the best way to assess statistical reasoning is through one-on-one interviews, as opposed to paper-and-pencil tests (e.g., Garfield, 2003). Upon completing the assessments, participants were asked to provide demographic information and then they were debriefed.

2.4. CODING

Conceptual knowledge of CIs We developed a measure of conceptual knowledge of CIs, and we used this measure as the primary outcome measure in our analyses.

Nine items from the conceptual knowledge assessment were taken to reflect conceptual knowledge of CIs. These items, listed in Table 2, were scored for correctness. Participants received one point for each item answered correctly. Items were scored by two independent coders, who obtained 91% agreement on scoring correctness. Disagreements were resolved via discussion.

Performance on the *correct* statements on the misconception assessment was also taken to reflect conceptual knowledge of CIs (Table 3). Participants whose average rating of the two correct statements was 4 or above received 1 point whereas participants whose average rating was below 4 received 0 points.

Composite scores for conceptual knowledge of CIs were calculated by summing the points earned on the nine items listed in Table 2, and the point earned for average ratings of the correct statements on the misconceptions assessment, for a maximum possible score of 10. This 10-item measure had acceptable reliability ($\alpha = 0.70$).

Sample mean understanding We also developed a composite score for sample mean understanding. This measure was used as a predictor in our analyses.

One item on the conceptual knowledge assessment was taken to reflect understanding of sample mean concepts ("Describe the difference between a sample mean and a population mean"). All students provided correct answers for this item, so our coding of this item focused on the depth of conceptual knowledge displayed in the responses. Participants who gave procedural definitions for each term (i.e., sample mean = sum of sample scores/n; population mean = sum of population scores/N) were classified as demonstrating *procedural* understanding. Responses that went beyond formulas and explained the relationship between a sample and population (i.e., responses that stated that a sample is a subset of a population) were classified as demonstrating *relational* understanding. Finally, participants who also mentioned the purpose of sampling (i.e., to make inferences about the population) were classified as demonstrating *generalizable* understanding. Agreement between coders was 88% for this item. Participants who displayed a generalizable understanding received one point for this item, and all other responses received zero points.

Two items on the misconception assessment were designed to assess the sample mean misconception (Table 3). Participants whose average rating of the two sample mean misconception statements was higher than 3 (reflecting endorsement of the misconception) received 0 points for this item, whereas participants whose average rating was 3 or below received 1 point.

Scores for sample mean understanding were created by summing the points earned on the sample/population mean item from the conceptual knowledge assessment and on the sample mean misconception statements on the misconceptions instrument. Thus, participants could receive composite scores for sample mean understanding from 0–2.

Estimation Responses to the open-ended items on the conceptual knowledge assessment were coded for references to estimation concepts. The six open-response and explanation items used in the conceptual knowledge of CIs measure (Table 2) were coded for this purpose. One additional open-response item was also coded for this purpose ("What are the advantages and disadvantages of 90% confidence intervals, relative to 99% confidence intervals?"). This item was not included in the conceptual knowledge of CIs measure, because participants interpreted it in various ways.

Participants received a score indicating the number of items, out of a possible seven, on which they mentioned estimation. Reliability was 94% for coding mentions of estimation.

Null Hypothesis Significance Testing (NHST) Responses to the open-ended items on the conceptual knowledge assessment were also coded for references to NHST concepts. The items coded for this purpose were the same as those coded for mentions of estimation, with one exception—the open-response item from the CIs assessment about the NHST interpretation of CIs was omitted because it specifically mentioned null hypothesis testing.

Preliminary analyses suggested that references to NHST were highly accurate overall, even when they occurred in the context of a longer incorrect response. Therefore, individual NHST statements were not scored for accuracy. Participants received a score indicating the number of items, out of a possible six, on which they mentioned NHST. Reliability was 94% for coding mentions of NHST.

Misconceptions Participants' ratings of the misconception statements on the misconception assessment were used as a metric for the existence of that misconception (see Table 1 for descriptions of the misconceptions). For misconceptions that were espoused in more than one statement (i.e., Sample Mean, Individual Scores; Table 3), the ratings of the statements were averaged to obtain one score for that misconception.

Responses to the open-ended items on the conceptual knowledge assessment were also coded for specific misconceptions. The items coded for this purpose were the same seven items that were coded for mentions of estimation. For each misconception, participants were given a score of 1 if they ever endorsed the misconception and a score of 0 if they either (a) never mentioned the misconception or (b) mentioned it as being incorrect. Reliability was 97% for coding misconceptions.

3. RESULTS

3.1. TEST TYPE

Participants who completed the assessment in paper-and-pencil format did not differ from those who completed the assessment while talking aloud, either in conceptual knowledge of CIs or in sample mean understanding. Additionally, talk-aloud participants were no more likely than paper-and-pencil participants to mention estimation, NHST, or any of the misconceptions in their responses to the open-ended items. The two groups of participants also did not differ in their ratings of the misconceptions. In light of this, data from all participants was aggregated for analysis.

3.2. CONCEPTUAL KNOWLEDGE OF CIs

Overall performance was mediocre, with composite scores for conceptual knowledge of CIs averaging 5.97 out of a possible 10 (range 1–10, 95% CI [5.18, 6.76]). We analyzed conceptual knowledge scores using a general linear model, and found that, on average, graduate students performed 1.75 points better than undergraduate students (graduate M = 6.89, SD = 2.58, range = 2–10; undergraduate M = 5.14, SD = 2.13, range = 1–8), $F(1, 38) = 5.53$, $p = 0.024$, $\eta_p^2 = 0.127$, 95% CI for difference in means [0.24, 3.26].

3.3. MISCONCEPTIONS

Considering each misconception separately, we examined the proportion of participants who mentioned the misconception and the ratings of the misconception items (Table 4). Because none of the rating variables met the normality assumption, we present only descriptive results for the combined sample. To evaluate group differences in the misconception ratings, we used bias-corrected and accelerated bootstrapping with 1,000 replications to estimate 95% confidence intervals for the differences in the mean ratings.

Sample mean (SM) Explicit statements indicative of the belief that a CI allows one to estimate the sample mean were rare, with only six participants (2 undergraduates and 4 graduates) ever displaying this misconception. As a representative example, in response to the item "Explain how to interpret the following confidence interval: 95% confidence interval $8.5 < \mu < 11.5$," one participant wrote, "We can say with 95% confidence that our sample mean falls within 8.5 units less than the mean or 11.5 units greater than the mean." The SM statements in the misconceptions assessment were rated as moderately accurate (M = 3.26, SD = 1.09). These ratings were included in the sample mean understanding measure, discussed below. Undergraduate students' average ratings of the SM statements were 0.35 points higher than graduate students' average ratings (undergraduate *M* = 3.43, *SD* = 1.12; graduate *M* = 3.08, *SD* = 1.04), 95% CI for mean difference [–0.95, 0.33].

Table 4. Proportion of participants who mentioned a particular misconception and average ratings on the misconception assessment for each group

Misconception	Undergraduate		Graduate	
	Mentioned Proportion (SE)	Rating Mean (SD)	Mentioned Proportion (SE)	Rating Mean (SD)
Sample Mean (SM)	9.5% (0.06)	3.43 (1.12)	21.0% (0.09)	3.08 (1.04)
Confidence Level (CL)	N/A	3.71 (1.38)	42.1% (0.11)	3.79 (1.36)
Individual Scores (IS)	14.2% (0.07)	1.80 (0.94)	N/A	1.32 (0.58)
Fixed Interval (FI)	42.8% (0.10)	4.19 (1.33)	68.4% (0.10)	3.00 (1.70)
Equality (EQ)	N/A	1.65 (0.99)	N/A	1.26 (0.73)

Confidence Level (CL) Eight participants made at least one statement expressing the CL misconception: the belief that the percent confidence associated with an interval indicates the percentage of replication means that will fall within the original interval. For example, in response to the item "Explain how to interpret the following confidence interval: 95% confidence interval 8.5 < μ < 11.5," one participant stated that " … a sample of whatever size that we got this confidence interval from, if we were to draw samples of that size, 95 times out of 100, the mean of that sample would lie between those two values." All eight of the participants who made such statements were graduate students. On the misconceptions assessment, the CL statement was rated as being fairly accurate (*M* = 3.75, *SD* = 1.35). Graduate students' average ratings of the CL statement were 0.08 points higher than undergraduates' average ratings (graduate *M* = 3.79, *SD* = 1.36; undergraduate *M* = 3.71, *SD* = 1.38), 95% CI for difference in means [–0.74, 0.90].

Individual Scores (IS) The IS misconception was the second least frequent in participant responses, with only three participants ever asserting that a CI provides information about individual scores. For example, when asked to define the term confidence interval, one participant wrote, "A confidence interval is an interval of whichever % you choose that between a range of [*x*, *y*] that % of the population will fall inside those scores."

Ratings of the IS statements were also very low (*M* = 1.56, *SD* = 0.81). Undergraduates' average ratings of the IS statements were 0.48 points higher than graduate students' average ratings (undergraduate *M* = 1.8, *SD* = 0.94; graduate *M* = 1.32, *SD* = 0.58), 95% CI for difference in means [–0.99, 0.019]. Two rating items assessed this misconception and both received low ratings (*M* = 1.35; *M* = 1.77). These low ratings align with other data suggesting that the IS misconception is relatively rare. For example, Fidler (2006) reported that, on two different items, only 8% and 11% of a sample of Australian first- and second-year psychology and ecology students endorsed this misconception.

Fixed Interval (FI) Use of language associated with the FI misconception was quite common, with about half of all participants (22 of 40) making a statement about the mean "falling" inside the interval. Ratings of the FI statement also indicated that participants generally believed it to be accurate (*M* = 3.62, *SD* = 1.61). Undergraduates' average rating of the FI item was 1.19 points higher than graduate students' average rating (undergraduate *M* = 4.19, *SD* = 1.33; graduate *M* = 3.00, *SD* = 1.70), 95% CI for difference in means [–2.03, –0.188].

Equality (EQ) No participant ever explicitly mentioned the idea that a CI indicates the likelihood that the sample mean equals the population mean. Ratings of the EQ statement were uniformly low (*M* = 1.46, *SD* = 0.88). Undergraduates' average ratings were 0.39 points higher than graduate students' average ratings (undergraduate *M* = 1.65, *SD* = 0.99; graduate *M* = 1.26, *SD* = 0.73), 95% CI for difference in means [–0.94, 0.21]. This was the only statement that did not receive a rating of 5 from at least one participant.

3.4. PREDICTOR MEASURES

We used general linear models to analyze whether values of each predictor measure varied as a function of participant group (i.e., undergraduate or graduate students). Each variable met the general linear model assumption of normality.

Sample mean understanding Overall, participants scored fairly low on sample mean understanding (M = 0.80 out of 2; 95% CI [.56, 1.03]). This low level of understanding aligns with other data indicating that many students have weak understanding of the sample mean and its relation to CIs (Fidler, 2006). Scores for undergraduate students and graduate students were similar (undergraduate M = 0.62, SD = 0.67; graduate M = 1.00, SD = 0.75), b_1 = 0.38, F(1, 38) = 2.90, p = 0.09, 95% CI for mean difference [–0.07, 0.83]. As predicted, understanding of the sample mean was positively related to conceptual knowledge of CIs, b_1 = 1.26, F(1, 38) = 5.95, p = 0.019, η_p^2 = 0.135, 95% CI for mean difference [0.215, 2.31].

Null Hypothesis Significance Testing (NHST) References to NHST were fairly common, with 15 participants mentioning NHST at least once. For example, one participant stated that a confidence interval is “the interval where your, um, where your calculated critical value should fall and then if the value of your null hypothesis falls within there, then you can’t reject it.” Undergraduate and graduate students mentioned NHST at similar rates (undergraduate M = 1.33, SD = 1.49 vs. graduate M = 1.16, SD = 1.12), b_1 = –0.17, F(1, 38) = 0.174, p = 0.679, 95% CI for mean difference [–1.02, 0.67]. Furthermore, although the relationship between mentioning NHST and conceptual knowledge was in the predicted negative direction, referencing NHST did not significantly predict conceptual knowledge of CIs, b_1 = –0.52, F(1, 38) = 3.24, p = 0.079, 95% CI for difference in means [–1.12, 0.065].

Estimation References to estimation were produced by about one-quarter (13 of 40) of participants. For example, when asked to explain the effect of increasing sample size on confidence level (all other factors remaining constant), one participant wrote, “Your estimate with a smaller sample is less precise/representative/etc., meaning it’s more likely your sample mean deviates from the population mean, widening the CI.” On average, graduate students produced 0.96 more mentions of estimation in their responses than undergraduate students (graduate M = 1.16, SD = 1.57; undergraduate M = 0.19, SD = 0.40), F(1, 38) = 7.42, p = 0.009, η_p^2 = 0.163, 95% CI for difference in means [0.248, 1.68]. In addition, as predicted, mentions of estimation were positively related to conceptual knowledge of CIs, b_1 = 0.97, F(1, 38) = 11.04, p = 0.001, η_p^2 = 0.225, 95% CI for mean difference [0.38, 1.56].

3.5. PREDICTING CONCEPTUAL KNOWLEDGE OF CIs

We also used a general linear model to analyze sample mean understanding, references to NHST, and references to estimation as predictors of conceptual knowledge of CIs (Table 5). There was a significant effect of sample mean understanding, such that better understanding of the sample mean was associated with greater knowledge of CIs, b_{SM} = 1.00, F(1, 36) = 4.29, p = 0.045, η_p^2 = 0.107, 95% CI for mean difference [0.02, 1.99]. Additionally, there was a significant effect of estimation, such that more mentions of estimation were associated with greater knowledge of CIs, b_{EST} = 0.73, F(1, 36) = 5.59, p = 0.023, η_p^2 = 0.135, 95% CI for mean difference [0.10, 1.36]. Mentions of NHST was not a significant predictor of knowledge of CIs when the other predictors were included in the analysis,

Table 5. Prediction of conceptual knowledge of CIs from sample mean understanding, null hypothesis significance testing mentions, and estimation mentions

Predictor	b	Std. error	p-value
SM Understanding	1.00	0.485	0.045
NHST Mentions	–0.28	0.279	0.316
Estimation Mentions	0.736	0.311	0.023

Model R^2 = 0.31. SM = sample mean; NHST = Null hypothesis significance testing

$b_{\text{NHST}} = -0.28$, $F(1, 36) = 1.03$, $p = 0.316$, $\eta_p^2 = 0.028$, 95% CI for difference in means [–0.85, 1.36]. In total, the effect of sample mean understanding, NHST mentions, and estimation mentions on conceptual knowledge was large, in that these three predictors accounted for 31% of the variance in conceptual knowledge of CIs.

4. DISCUSSION

In this study, conceptual knowledge of CIs was assessed in both undergraduate and graduate psychology students. The current findings replicate previous work suggesting that CIs are challenging (e.g., Fidler, 2006), in that performance on the assessment was only mediocre, despite the fact that all participants had received instruction about CIs in at least one statistics class.

4.1. PERFORMANCE BY UNDERGRADUATE VS. GRADUATE STUDENTS

Unsurprisingly, graduate students outperformed undergraduates on many items. In fact, on all items for which the performance of the two groups differed substantially, the graduate students performed better. There was, however, one misconception that was espoused more frequently by graduate students. The confidence level (CL) misconception was mentioned by eight graduate students but was never mentioned by an undergraduate. Based on the content of the responses, it seems that espousing this misconception may actually reflect slightly better knowledge of CIs. Specifically, the CL misconception requires recognizing the relevance of replication (repeated sampling) to the definition of the confidence interval. Thus, although graduate students erred in characterizing the specific relationship between confidence level and replication—an error also commonly made by researchers (Cumming & Maillardet, 2006)—they demonstrated a slightly more sophisticated understanding than undergraduates, who rarely linked CIs and replication.

4.2. CONFIDENCE INTERVAL MISCONCEPTIONS

Previous work has identified a number of misconceptions that students and researchers have about CIs (Table 1). The current findings suggest that although some of these misconceptions are fairly common, others are less frequent. The individual scores (IS) misconception (i.e., a CI gives the range of individual scores) and the equality misconception (i.e., a CI tells the likelihood of the sample mean being equal to the population mean) were mentioned very rarely and were generally given low ratings. The sample mean misconception (i.e., a CI allows one to estimate the sample mean) was rarely mentioned, but it did receive moderately high ratings. The confidence level misconception (i.e., the confidence level of a calculated interval indicates the percentage of replication means that will fall within the original interval) was mentioned frequently by graduate students and was generally rated highly, suggesting that it may be a widely held belief. The fixed interval misconception (i.e., a CI is a fixed interval, within which a moving parameter may or may not fall) was also mentioned frequently, and this may be reflective of the frequent use of fixed-interval language in statistics education (see Grant & Nathan, 2008).

4.3. PREDICTING PERFORMANCE

Based on previous work, we hypothesized that mentioning NHST and estimation, the two main uses of CIs, would each relate to performance, but in different ways. The association of NHST mentions and conceptual knowledge of CIs was in the expected negative direction, but was not significant. As predicted, however, mentions of estimation were positively associated with conceptual knowledge of CIs. These findings suggest that it may be important to consider how the estimation aspect of CIs is taught in statistics classes.

One additional concept, sample mean understanding, was also hypothesized to be relevant to CI knowledge. As predicted, participants with greater understanding of the relationship between a sample mean and the population mean had higher levels of conceptual knowledge of CIs.

When NHST, estimation, and sample mean understanding were all included in a single model, estimation and sample mean understanding were the only two significant predictors of conceptual knowledge of CIs. Estimation and sample mean understanding may be useful to students because they highlight general principles that are important for deep understanding of statistics, more generally, and CIs, in particular.

4.4. LIMITATIONS

We created our assessment instruments to evaluate participants' knowledge of CIs and related concepts, and to diagnose potential misconceptions. Our assessments represent a first step in assessing knowledge of CIs and related concepts, and they go beyond evaluating participants' abilities to calculate CIs. We acknowledge, however, that our assessments could be improved in many ways.

One potential issue has to do with the response scale we used for the misconceptions assessment, which ranged from *very inaccurate* to *very accurate*. Participants may have found it confusing to judge the *degree* of accuracy of a statement, rather than simply to judge it as accurate or inaccurate. An alternative approach would be to ask participants to judge whether each item is correct or incorrect, and then to ask them to provide a rating of their confidence in that decision.

Another issue has to do with the item with which we intended to assess the Fixed Interval misconception on the misconceptions assessment (Table 1). Although we intended this item to indicate that the interval was fixed and the parameter could vary (falling into the interval or not), we realized after the data were collected that the wording was ambiguous, and that some participants may not have interpreted this item in the way we intended. This is potentially problematic because we included only one item to assess the Fixed Interval misconception. We recommend that future studies that seek to investigate the Fixed Interval misconception use a clearer statement of this misconception.

The items that we used to assess the individual scores misconception could also be improved. This misconception is sometimes expressed as "95% of the data are included in the confidence interval" or "the confidence interval shows the data within one standard deviation of the mean" (see e.g., Castro Sotos et al., 2007; Fidler, 2006; Grant & Nathan, 2008)—ideas that were not reflected in the specific items we used. Future studies could include a wider range of misconception items and could examine the correlations among them.

Another set of limitations has to do with our sample. Both participant groups were small (21 undergraduates and 19 graduate students) and fairly homogeneous, and the groups did not represent random samples of psychology students. All of the undergraduate students had taken or were taking the same undergraduate statistics course, and all of the graduate students were from the same department and had taken the same graduate statistics course, although not all in the same year. Thus, one should be cautious about generalizing the findings to psychology undergraduates or graduate students at other universities, who might use different textbooks or who might have different patterns of course work. Indeed, it is possible that some of the misconceptions that we observed may have been inadvertently reinforced by the textbooks that students used or by the instruction or curricular sequencing that they encountered in their statistics courses (see Grant & Nathan, 2008).

In light of these limitations, we suggest that future studies include larger samples of students and examine how curricular materials, such as textbooks, influence students' knowledge of CIs. Future studies should also focus more directly on participants' interpretations of CIs and on how context affects the knowledge about CIs that participants activate and draw on. Future studies might also include additional items to directly assess knowledge of sample mean, estimation, and NHST.

Although our focus in this paper has been on confidence intervals construed within a frequentist perspective, an alternative would be to employ a Bayesian framework, which assumes that a parameter has a probability distribution. Within the Bayesian framework, a 95% posterior probability, or credible, interval means that the probability that the parameter lies in the interval is 95%. This Bayesian interpretation—though not accurate for confidence intervals—is highly intuitive (see Gurrin, Kurinczuk, & Burton, 2000), and indeed, it is similar in spirit to two of the misconceptions identified within the frequentist perspective, the confidence level and fixed interval misconceptions.

4.5. CONCLUSION

The current findings replicate previous work suggesting that CIs involve difficult statistical concepts. Further, they document the existence of some of the previously identified misconceptions in participant groups that have not previously been studied. Although three foundational constructs were hypothesized to influence conceptual knowledge of CIs, only two—estimation and sample mean understanding—emerged as significant contributors to conceptual knowledge of CIs. Previous work has focused on concepts that might be detrimental to CI understanding, but the present work suggests that there may be some concepts, such as estimation, that are actually beneficial.

The current findings highlight the need for additional research regarding ways to improve lessons about CIs in order to promote deep conceptual understanding and prevent misconceptions. Based on this work, it appears that lessons that focus on the relationship between samples and populations and on the use of CIs for estimation might be particularly effective.

ACKNOWLEDGEMENTS

The research reported here was supported by the U.S. Department of Education, Institute of Education Sciences, through Grant #R305B090009 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the Institute nor the U.S. Department of Education. The authors thank Jaclyn Magee and Chloe VanDerWeele for their help with data collection and coding.

REFERENCES

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Society (2017). *Submission guidelines*.
[Online: https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STAT]

American Statistical Association. (2016). *Statement on statistical significance and p-values* [Press release].
[Online: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf]

Bisson, M. J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, *2*(2), 141–164.

Bude, L., Imbos, T., van de Wiel, M. W. J., & Berger, M. P. (2011). The effect of distributed practice on students' conceptual understanding of statistics. *Higher Education*, *62*(1), 69–79.

Bude, L., van de Wiel, M. W. J., Imbos, T., & Berger, M. P. F. (2010). The effect of directive tutor guidance on students' conceptual understanding of statistics in problem-based learning. *British Journal of Educational Psychology*, *81*(2), 309–324.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 26.

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge of mathematics. *Developmental Review, 34*(4), 344–377.

Cumming, G. (2006). Understanding replication: Confidence intervals, *p*-values, and what's likely to happen next time. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7),* Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/7D3_CUMM.pdf]

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie [Journal of Psychology]*, *217*(1), 15–26.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*, 138–146.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*(3), 230–232.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11*(3), 217–227.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and error bars. *Understanding Statistics, 3*(4), 299–311.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 62*(2), 28–58.
[Online: https://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* (Unpublished doctoral dissertation). University of Melbourne, Australia.

Fidler, F. (2006). Should psychology abandon *p*-values and teach CIs instead? Evidence-based reforms in statistics education. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7)*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots7/5E4_FIDL.pdf]

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), 136–143.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p*-values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie/Journal of Psychology, 217*(1), 27–37.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15*(2), 119–126.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1–13). Amsterdam: IOS Press.

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The interpretation of scholars' interpretations of confidence intervals: Criticism, replication, and extension of Hoekstra et al. (2014). *Frontiers in Psychology, 7,* 1–12.

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

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

Graham, A. T., & Thomas, M. O. J. (2005). Representational versatility in learning statistics. *International Journal for Technology in Mathematics Education, 12*(1), 3–4.

Grant, T. S., & Nathan, M. (2008). Students' conceptual metaphors influence their statistical reasoning about confidence intervals. WCER Working Paper No. 2008-5. Wisconsin: Wisconsin Center for Education Research.
[Online: https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2008_05.pdf]

Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the behavioral sciences* (9th ed.). Belmont, CA: Wadsworth/Cengage Learning.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman D. G. (2016). Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350.

Gurrin, L. C., Kurinczuk, J. J., & Burton, P. R. (2000). Bayesian statistics in medical research: An intuitive alternative to conventional data analysis. *Journal of Evaluation in Clinical Practice*, *6*(2), 193–204.

Henriques, A. (2016). Students' difficulties in understanding of confidence intervals. In D. Ben-Zvi & K. Makar (Eds.), *The Teaching and Learning of Statistics* (pp. 129–138). Cham, Switzerland: Springer.

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*, 137.

Hoekstra, R., Morey, R. D., Rouder, J. N., Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164.

Holte, J. M. (2003, April). *Teaching confidence intervals*. Paper presented at the Spring Meeting of the Mathematical Association of America, North Central Section, St. Paul, Minnesota.
[Online: https://pdfs.semanticscholar.org/83e4/bd866088792e42c8a83353095a03ff9dc8e68.pdf]

Hong, E., & O'Neil Jr., H. F. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology, 84*(2), 150–159.

Jones, K. A., Jones, J. L., & Vermette, P. J. (2011). Putting cognitive science behind a statistics teacher's intuition. *Teaching Statistics, 33*(3), 85–90.

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Orlando, FL: Harcourt Brace College Publishers.

Leppink, J., Broers, N. J., Imbos, T., van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Self-explanation in the domain of statistics: An expertise reversal effect. *Higher Education,63*, 771–785.

Lockwood, E., Yeo, A., Crooks, N. M., Nathan, M. J., & Alibali, M. W. (2014). Teaching about confidence intervals: How instructors connect ideas using speech and gesture. In W. Penuel, S. A. Jurow, & K. O'Connor (Eds.), *Learning and becoming in practice: Proceedings of the Eleventh International Conference of the Learning Sciences*. Boulder, CO: University of Colorado.

Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal*, *1*(2), 22–37.
[Online: https://iase-web.org/documents/SERJ/SERJ1(2).pdf]

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123–137.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*, 429–434.

Pfannkuch, M., Wild, C. J., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM–International Journal on Mathematics Education*, *44*(7), 899–911.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth: A course in methods of data analysis.* Pacific Grove, CA: Duxbury/Wadsworth Group/Thomson Learning.

Reaburn, R. (2014). Students' understanding of confidence intervals. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: iase-web.org/Conference_Proceedings.php?p=ICOTS_9_2014]

Smith III, J. P., DiSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163.

MARTHA W. ALIBALI
Department of Psychology, University of Wisconsin – Madison
1202 W. Johnson St.
Madison, WI 53706 USA

APPENDIX A: ITEMS ON THE CONCEPTUAL KNOWLEDGE ASSESSMENT IN THE ORDER PRESENTED

1. Define the term “confidence interval.”

Explain how each of the following affects the width of a confidence interval:

2. Increasing the sample size
3. Increasing the sample variability
4. Increasing the level of confidence (the percent confidence)

5. What are the advantages and disadvantages of 90% confidence intervals, relative to 99% confidence intervals?
6. In a study of the effects of marijuana use during pregnancy, measurements on babies of mothers who used marijuana during pregnancy were compared to measurements on babies of mothers who did not. A 95% confidence interval for the difference in mean head circumference (non-use minus use) was 0.61 to 1.19 cm. What can be said from this statement about the hypothesis that the mean difference is zero?
7. True or False: If all other factors are held constant, an 80% confidence interval is wider than a 90% confidence interval. Please explain your choice.
8. True or False: If all other factors are held constant, a confidence interval computed from a sample of n = 25 is wider than a confidence interval computed from a sample of n = 100. Please explain your choice.
9. True or False: If all other factors are held constant, a confidence interval computed from a sample with high variability is narrower than a confidence interval computed from a sample with low variability. Please explain your choice.
10. Imagine you are describing confidence intervals to a beginning statistics student. Explain how to interpret the following confidence interval: 95% confidence interval $8.5 < \mu < 11.5$.
11. Write the formula for a confidence interval. Label all pieces of the equation.
12. Describe the difference between a sample mean and a population mean.

APPENDIX B: ITEMS ON THE MISCONCEPTION ASSESSMENT IN THE ORDER PRESENTED

Rate the following statements on a scale from 1 (Very Inaccurate) to 5 (Very Accurate). Circle your response.

a. A confidence interval gives you the range of possible values for the sample mean.
b. If you were to conduct an infinite number of experiments exactly like the original experiment, a 95% confidence interval would contain 95% of the sample means from these experiments.
c. A confidence interval gives you the range of the individual scores.
d. A 95% confidence interval is the interval for which you are 95% certain that it contains the population mean.
e. If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean.
f. A confidence interval gives you the range of the individual scores within one standard deviation of the population mean.
g. A 95% confidence interval is the interval for which you are 95% confident that the population mean falls within it.
h. A 95% confidence interval indicates that there is a 95% chance that the sample mean equals the population mean.