WCER Working Paper No. 2008-5
August 2008

Students' Conceptual Metaphors Influence Their Statistical Reasoning About Confidence Intervals

Timothy S. Grant
Department of Educational Psychology
University of Wisconsin–Madison
tsgrant@wisc.edu

Mitchell J. Nathan
Department of Educational Psychology/
Wisconsin Center for Education Research
University of Wisconsin–Madison
mnathan@wisc.edu

WCER
Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

Grant, T. S., & Nathan, M. J. (2008). *Students' conceptual metaphors influence their statistical reasoning about confidence intervals* (WCER Working Paper No. 2008-5). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved [e.g., August 15, 2008,] from http://www.wcer.wisc.edu/publications/workingPapers/papers.php

The research reported in this paper was supported by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Students' Conceptual Metaphors Influence Their Statistical Reasoning About Confidence Intervals

Timothy S. Grant and Mitchell J. Nathan

Objectives

Confidence intervals are beginning to play an increasing role in the reporting of research findings within the social and behavioral sciences (American Psychological Association, 2001; Garfield & Ben-Zvi, in press) and, consequently, are becoming more prevalent in beginning classes in statistics and research methods. In delineating the American Statistical Association guidelines for what it means to be statistically educated at the college level, Franklin and Garfield (2006) identified understanding and proper use of confidence intervals as part of the basic knowledge of statistical inference.

As a reminder, for some significance level set at α, a *confidence interval* is a contiguous set of points for which we are 1-α % confident that one of the points in the set is equal to the fixed, but often unknown, population parameter or mean. For example, if as a result of one experimental sample we get a 95% confidence interval for the population mean μ, the interval might look like

$$8.5 < \mu < 11.5 .$$

If the experiment is repeated and new samples are taken from the general population, the population mean will remain unchanged, while the distance between the limits and the center of the interval will change to reflect the properties of each new sample.

Confidence intervals are gaining increasing importance in statistics education (Garfield & Ben-Zvi, in press). In their "Top Ten" list of recommendations for teaching the reasoning of statistical inference, Rossman and Chance (1999) made "Accompany tests of significance with confidence intervals whenever possible" #7 on the list. Reasons cited for the rise in the use of confidence intervals include the fact that they provide more information—including the sample mean and sample variance, which are estimates of the population mean and population variance—than the traditional inferences drawn from hypothesis testing. The sample mean is present in the interval as the center, and the width of the interval is a function of the sample variance and sample size. Furthermore, a simple test at a significance level of α of whether the parameter μ is equal to some specific value μ_0 is simply to check if μ_0 falls inside the interval.

This shift in the uses of statistics becomes problematic, however, if the confidence interval is misinterpreted. Some researchers (Belia, Fidler, Williams, & Cumming, 2005; Schenker & Gentleman, 2001) have pointed out that comparisons are often made between a confidence interval and a point estimate without noting the error associated with the point estimate. Furthermore, when presented with multiple confidence intervals, as in an analysis of variance, students may fail to properly combine the information contained in the two intervals, leading to hypothesis tests that do not have the proper level of significance.

In a national sample of 817 students in introductory statistics classes at 25 institutions of higher education across 18 states, the **C**omprehensive **A**ssessment of **O**utcomes in a First **S**tatistics Course (CAOS; delMas, Garfield, & Chance, in press) yielded mixed results on gains in conceptual understanding of confidence intervals. On the positive side, posttests showed a significant increase over pretests in the percentage of students correctly identifying an invalid interpretation of a confidence level (that for α of 5%, the 95% refers to the percentage of population data values between the confidence limits). The data also showed significant gains in students' abilities to provide the standard interpretation of a confidence interval as a set of plausible values of the unknown population parameter (μ), based on a random sample taken from that population. However, students showed poor understanding of, and no measurable gains in their ability to detect, two misinterpretations of a confidence level—namely, (a) that 95% represents the percentage of the sample data that lies between the two confidence limits; and (b) that 95% is the percentage of all of the possible sample means between the upper and lower limits of the confidence interval.

Garfield, delMas, and Chance (n.d.) identified some of the most common misconceptions about confidence intervals as:

- There is a 95% chance the confidence interval includes the sample mean.
- There is a 95% chance the population mean will be between the two values (upper and lower limits).
- Ninety-five percent of the data are included in the confidence interval.
- A wider confidence interval means less confidence.
- A narrower confidence interval is always better (regardless of confidence level).

Objectives

While earlier studies have demonstrated misconceptions and misapplications of confidence intervals, the research literature on statistics education reveals little about the conceptual underpinnings of these misconceptions. The objectives of this study are to (a) show that the theory of conceptual metaphor as delineated in contemporary embodied cognition is a useful framework for describing statistics students' conceptions of confidence intervals; and (b) provide empirical evidence from discourse and gesture that graduate students in social science use at least two competing conceptual metaphors for confidence limits that have important implications for the understanding and application of statistics and for the reform of statistics education.

Theoretical Framework

Statistical Background

Confidence intervals are historically a product of estimation theory (Stigler, 1986). Originally confidence intervals were used to bound the distance between an estimate and the unknown but fixed population parameter. As such, the confidence interval for a mean would be computed as

$$\left|\mu - \bar{x}\right| < d \ ,$$

where *d* is the bound of the estimate. Eventually, the distance is measured in standard deviations (σ), and probability statements can then be made as in the following expression:

$$\bar{x} - Z_{\alpha/2}\, \sigma\!\Big/\!\sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}\, \sigma\!\Big/\!\sqrt{n} \ .$$

The confidence interval is considered a set of estimates for μ. This interval is one of many that could be computed by repeatedly sampling the population and computing intervals. If this process is repeated many times, 1-α of the intervals will be correct in the sense that they will contain the true parameter. The key is that from sample to sample, the interval will change, both in the range between upper and lower bounds and in its central location (Mills, 2002). The above relation assumes that the variance (σ^2) is known, and so the width only changes with sample size (*n*). In general, the variance is unknown, and the above expression is applied with an estimated variance and has a Student's *t*-distribution. In this latter case, the width is more volatile, but for simplicity, we will consider the case in which the variance is known.

In hypothesis testing, by contrast, boundaries are set by the mathematical question being asked and a sample statistic falling into this region. Simply stated, a rejection region and its complement, the acceptance region, are calculated based on the variance and the hypothesized value of the parameter μ_0. Thus, the acceptance region is calculated as

$$\left(\mu_0 - Z_{\alpha/2}\, \sigma\!\Big/\!\sqrt{n}, \mu_0 + Z_{\alpha/2}\, \sigma\!\Big/\!\sqrt{n}\right).$$

Comparing this expression to the one above shows that the similarities are quite strong, which can cause confusion (Thompson, Saldanha, & Liu, 2004). If the confidence interval is written as a set in the same way the acceptance region is written above, the only difference is that the sample mean, μ, is replaced with a hypothesized mean, μ_0. At this point, it should also be noted that the regions in hypothesis testing are generally talked about in terms of the rejection region; *acceptance region* is a misnomer since the null hypothesis is never actually accepted (although it is possible for it not to be rejected). If one conducts many hypothesis tests based on different samples taken from a single population, this acceptance region would be constant for each test. For each sample, one would calculate a new sample mean and see if it fell inside or outside this interval.

Conceptual Metaphors

Embodied cognition is an emerging epistemological framework that examines mental behavior in relation to (a) the physical and social environment within which people operate and (b) the perception- and action-based systems of the body (Glenberg, 1997; Nathan, in press). By casting cognition explicitly in terms of interactions between agents and the world rather than as isolated and amodal symbolic computation, embodied cognition reframes some of the central issues of the study of thought and behavior.

Theories of embodied cognition hypothesize, for example, that mathematical ideas can be explained as a system of conceptual metaphors for events and objects in the world (Lakoff & Nunez, 2000). *Conceptual metaphor* is a cognitive mechanism that allows people to reason about some new kind of abstract thing (e.g., a mathematical entity, like numbers) as if it were a familiar thing (like a collection of objects). Critically, a metaphoric mapping is both inference preserving and grounded. It is *inference preserving* in that it maintains the conceptual structure of the familiar (or *source*) domain and applies it to the new (*target*) domain. Thus, we can conceive of combining numbers—even numbers we have never encountered before—in the same way we combine sets of objects, and we can expect that their aggregation will equal their arithmetic sum. The *grounded* quality of metaphoric mappings means that the source domain of the mapping is ultimately connected to states of the world, as filtered through the neural system. That is, to check our ideas about addition, we can feel and see how the aggregation of collections of objects combine to form a collection that is the size of the combined sets.

Conceptual Metaphors for Confidence Intervals

It is within the context of this theory of conceptual metaphors that we describe the following two metaphors for confidence intervals: (a) Confidence Intervals Are Changing Rings Around a Fixed Point (*Changing Ring metaphor*); and (b) Confidence Intervals Are Changing Points on a Fixed Disk (*Fixed Disk metaphor*). In the Changing Ring metaphor (Table 1), confidence intervals are moving disks of various diameters covering a fixed but unknown point. This is like pitching horseshoes of varying widths to encircle a stake fixed in the ground. Key to this correct conceptual metaphor is representing the many ways in which the critical values of the confidence interval change, since the confidence interval is a property of a sample but not necessarily of the larger population from which the sample was taken. As such, the diameter of the disk, which maps to the length of the confidence interval, changes from sample to sample as the size and standard deviation of the sample change. Similarly, the location of the center of the interval (disk) changes and is determined by the sample mean. This is contrasted to the population parameter, or population mean, which is fixed across samples but generally unknown.

Table 1
Changing Ring Metaphor

Source domain **Changing ring around a fixed point**		***Target domain*** **Confidence interval**
Fixed point (stake in the ground)	→	True value of parameter
Changing ring	→	Confidence interval
Ring diameter	→	Size of interval
Ring center	→	Estimate of parameter
Ring falls on fixed point	→	Correct interval (One that contains true parameter)

Source domain **Changing ring around a fixed point**		***Target domain*** **Confidence interval**
Ring does not fall on fixed point	→	Incorrect interval
A particular ring	→	Random sample
95% of rings fall on fixed point	→	95% confidence interval

In general, the boundaries of the confidence interval are properties of the sample, and for any particular population parameter there are infinitely many confidence intervals of varying sizes and central tendencies that have the same probability (calculated as 1-α, for a reliability set at α, typically 5%) of covering the true population parameter. Once an interval is determined, the probability the interval covers the true parameter is exactly 0 or 1 (i.e., it either does or does not include the population mean), though the value is generally unknown. The probability statement of a confidence interval is that if many intervals are calculated at a level of α, 1-α percent of the intervals will be correct and contain the true parameter with probability 1.0 (Wardrop, 1995).

The Fixed Disk metaphor (Table 2) conceptualizes confidence intervals as fixed-diameter disks onto which successive points are placed. This metaphor is similar to the idea of throwing darts at a dartboard where the board is a fixed ring and the darts represent the various points. In this metaphor, the belief is that the population parameter can change from sample to sample, which contradicts an essential assumption of inferential statistics. The interval in this metaphor is taken to be of fixed length, and each experiment results in placing a new parameter onto the fixed-diameter disk.

Table 2
Fixed Disk Metaphor

Source domain **Changing point on a fixed disk**		***Target domain*** **Confidence interval**
Fixed disk	→	Confidence interval
Changing point	→	Population parameter
Disk diameter	→	Size of interval
Disk center	→	Estimate of parameter
Point falls on fixed disk	→	Correct interval (One that contains true parameter)
Point does not fall on fixed disk	→	Incorrect interval
A particular point	→	Random sample
95% of points fall on fixed disk	→	95% confidence interval

One reason we suggest this metaphor is a suspected confusion between acceptance regions in hypothesis testing and confidence intervals, as described above. As a conceptual metaphor, hypothesis testing would be defined as in Table 3. The similarities of Table 3 and

Table 2 might lead to confusion between hypothesis testing and confidence intervals. Such confusion will generally lead to a misinterpretation of the confidence interval.

Table 3
Hypothesis Testing Is a Changing Point on a Fixed Disk Metaphor

Source domain **Changing point on a fixed disk**		***Target domain*** **Hypothesis testing**
Fixed disk	→	Acceptance region
Changing point	→	Test statistic
Disk diameter	→	Size of acceptance region (Level of test)
Disk center	→	Hypothetical value
Point falls on fixed disk	→	Fail to reject null hypothesis
Point does not fall on disk	→	Reject null hypothesis
A particular point	→	Sample test statistic
About 95 points out of 100	→	Level .05 test

Speech and Gesture as a Window into the Mind

It is believed that gestures and speech can reveal students' mental states and provide insights into their understanding and metaphors of mathematical concepts (e.g., Alibali, Bassok, Solomon, Syc, & Goldin-Meadow, 1999; Goldin-Meadow, 2003; McNeill, 1992; Nathan & Bieda, 2006). Specifically of interest in this work on confidence intervals are gestures and speech that indicate whether (a) the critical values of the interval are fixed or variable points and (b) the population parameters are stable or changing across samples. Gestures play a particularly valuable role in the assessment of student understanding of statistics, since much of the content of statistics is graphical and spatially organized.

Method

In an initial effort to assess the understanding of confidence intervals exhibited by students in a graduate course on statistical methods for the social sciences (a second course in a 3-semester sequence), we interviewed and videotaped 3 female volunteers who scored above average in the class. In addition to hypothesis testing and confidence intervals, the course curriculum addressed descriptive statistics, discrete distributions, the standard normal and Student's *t*-distributions, tests of association, ANOVA, and chi-square and *F*-distributions.

Coding System for Students' Responses

We asked students several conceptual and calculation-based problems (see Appendix). The interviews were videotaped, and students' responses coded for evidence in speech and gesture of either the Changing Ring (CR) or the Fixed Disk (FD) metaphor, using the criteria presented in Table 4.

Table 4
Coding Criteria

Code	Criterion
1	[CR] References to intervals that have boundary points that change in both location and width from experiment to experiment
2	[CR] References to a fixed population parameter
3	[CR] References to the true mean or population mean being unknown
4	[CR] Statements and gestures drawing analogies referring to estimation
5	[CR] References to the center of an interval as the sample mean
6	[FD] References to a population parameter being free to change
7	[FD] References to fixed intervals or fixed critical values, including hand gestures and verbal statements suggesting firm boundaries and dimensions that do not vary or change
8	[FD] Statements and gestures drawing analogies to acceptance regions or hypothesis testing
9	[FD] Statements that indicate that the width of the interval is the sole factor in determining the probability of a confidence interval
10	[FD] Statements that the center of the interval is the population parameter μ
11	[Both CR and FD] References to intervals that change in location, shifting in position, but maintain a constant width
12	[Neither] References to the population parameter falling inside a confidence interval (technically incorrect)

These coding criteria are based primarily on what is dynamic and what is static in each of the two competing metaphors. For this reason, Codes 1–5, which represent a fixed population parameter and intervals of varying sizes and locations, are examples of the Changing Ring metaphor. Since this methodology is an outgrowth of estimation theory, we coded statements referring to estimation as evidencing this metaphor. Finally, recognition that the true parameter is unknown was coded as evidence of this metaphor since it is a tacit acknowledgment that the center of the interval is not the population parameter.

Codes 6–10 are examples of the Fixed Disk metaphor. With this metaphor, there is often a tendency to believe the population parameter is not fixed; the statements and gestures represented in Codes 6–10 are examples of this way of thinking. References to the intervals' being fixed are also examples of the Fixed Disk metaphor. As the metaphor describes something similar to what does happen in hypothesis testing, references that draw the parallel or refer to acceptance or rejection regions were coded as evidence of the metaphor. Finally, references that draw a one-to-one relationship between the width of the interval and the confidence level, which ignores the effects of sample size and uncertainty in sample variation, are examples of the Fixed Disk metaphor.

The final two codes are more difficult to assess. Code 11 applies to references to an interval that changes location but whose width remains constant. We coded these references as evidence of both metaphors because the disk is changing, consistent with the Changing Ring metaphor, and the width is constant, consistent with the Fixed Disk metaphor. Finally, the Code 12 statement that "there is a 95% chance that the true parameter falls into the interval"—while conveying the Fixed Disk metaphor—is too common to be used as an indication of students' understanding and thus was not scored for either metaphor.

Examples of Codes from Students' Responses

For gestures that are dynamic, a single photograph is not enough to convey the message. For others, such as those conveying a fixed mean or a fixed interval, a single photograph may suffice. Figure 1 shows an example of a gesture indicating a fixed interval. The participant indicates two fixed points defining an interval by marking the boundaries a fixed distance apart with her index finger and pinkie. This type of gesture is an example of Code 7 and represents the incorrect metaphor. A similar type of gesture is possible to indicate the population parameter or mean, as shown in Figure 2. Instead of two points a fixed width apart, Figure 2 illustrates a marking of a single fixed point. This is consistent with the Changing Disk metaphor and is an example of Code 2. Figures 3 and 4 are examples of gestures that contrast fixed and moving boundaries. In Figure 3, the student's two open hands move outward to define a continuum (a set of many possible values). A variation on this theme is for the subject to rotate his or her hands in circles, again indicating the possibility of change. Figure 4 shows the student defining a specific interval by placing her two hands on the desk to mark the boundaries of the interval. Figure 4 is an example of fixed boundaries and thus is similar to Figure 1. If a student were to repeat the gesture in Figure 4, moving the location and width between hands in a dynamic fashion, the gesture would tend to represent the first criterion of a shifting interval with varying widths. Another example of how these gestures can be used dynamically is the raising and lowering of the gesture to indicate falling, such as repeatedly pointing to the mean as if it were falling.

Materials and Procedure

The interview questions (see Appendix) were primarily designed around a simple confidence interval for a population mean. The questions were given to each student in order during a one-on-one interview with the first author, and the responses were videotaped. Each participant was first presented with an example of a confidence interval. Question 1 then asked the participant to explain the confidence interval to a fictional student who was new to statistics. While the participant generally led the discussion, probes and follow-up questions were used to clarify which parts of the interval were seen by the participant as changing from experiment to experiment and to remind the participant to state explicitly an interpretation of the probability statement inherent in a 95% confidence interval.

In Question 2, the participant was provided with a set of sample statistics from a hypothetical experiment and asked to construct a confidence interval. The participant was then presented with a second set of hypothetical statistics based on the same population. The participant was asked to construct a second confidence interval and then to discuss the two confidence intervals. Finally, Question 3 presented two graphical depictions, each showing three “snapshots” leading to the placement of μ along a number line. The final image of μ lying between two interval boundaries was the same for each set of snapshots; only the steps leading up to the placement of μ changed. The first snapshot sequence showed μ as a fixed point along a number line, with two critical values placed to the right and the left of it. The second sequence showed two fixed critical values already located on the number line, with μ placed somewhere in the interval between them. Students were asked to indicate which of the two sequences was closer to their notion of finding a confidence interval and then to explain their reasoning. The correct interpretations of the confidence intervals were eventually presented to each participant following their responses so that no harm was done by this study.

The interviews were transcribed and coded based on the coding criteria above, with attention to the comments and gestures made. Since the sample size was small (n = 3) and the study obviously lacks power, statistical analysis of the frequency of codes was not conducted, though we show the data in Table 5. Instead, we used qualitative analysis to understand students’ conceptions of confidence intervals, develop empirically informed hypotheses about these conceptions, and inform future work.

Table 5
Reference to Correct and Incorrect Metaphors During Interviews

Participant	References to incorrect metaphor	References to correct metaphor	Proportion incorrect metaphor
1	12	8	0.600
2	9	6	0.600
3	7	16	0.304

Results

The coding of the think-aloud, videotaped interviews showed evidence that students were using the Fixed Disk metaphor for confidence intervals. This metaphor is incorrect as it creates the impression that (a) the confidence interval is a property of the population and fixed and (b) the population parameter is not fixed and actually moves within the interval. As shown by Table 5, two of the three students referred more often to the Fixed Disk metaphor than to the Changing Ring metaphor.

An interesting side note is that as students were encouraged by the interviewer to explain themselves further, they expressed more doubt and tended to refer to each of the two metaphors in turn. They generally concluded their first answer was correct and restated that metaphor. As such, the proportions of correct versus incorrect references tended to be around 1/3 and 2/3. An additional point of interest is that all three participants made reference to both metaphors. This indicates that while they were confused—possibly by the juxtaposition of the topics of hypothesis testing and confidence intervals or the similarities in calculating these ranges—both metaphors exist and influence students’ conceptualization of inferential statistics. It may be possible to improve statistics education by assisting students to differentiate the two topics and recognize that although the mechanics and notation are similar, the concepts are distinct.

The first participant responded to Question 1 by saying that a researcher is "95% confident that the actual population mean falls between these two numbers." The student gestured by marking two fixed points, consistent with Code 7 (Fixed Disk). When asked what the center of the interval was, the student responded, "I think this is the center, μ, the population mean," interpreted as an example of Code 10 (Fixed Disk). The student then pointed to the center of the confidence interval. Further evidence of the student's belief that the interval was static was the discussion of the interval's shifting from sample to sample. The student contrasted open spreading hands in general, saying "I'm thinking it is a continuum," but kept her closed hands a fixed distance apart when discussing intervals and how they "would shift along the continuum." Interestingly, when asked directly, the student was able to recognize that the width of the interval could change. Her discomfort with this fact was noticeable, however, in her response to Question 2 when confronted with two intervals for the same parameter that had different widths. She commented: "Seems weird, that there is a 95% chance it falls in here [points to larger interval] and a 95% chance it falls in here [points to smaller interval]." As this response demonstrates a belief that the width is the sole measure of confidence, it is an example of Code 9 (Fixed Disk) and is support for the incorrect metaphor. Basically, the student was uncomfortable with the disk changing as this was inconsistent with her metaphor. Since this student tended to prefer the Fixed Disk metaphor, one might predict the student would prefer the graphical representation of μ falling into a fixed interval for Question 3. This was indeed the case.

The responses of the second participant wandered between estimation, Z-scores, and acceptance regions. In responding to Question 1, the student described a confidence interval as a measure of how far the sample mean is from the true score. The description is of a single test score, a sample size of one, and wanting to know "how sure are we that this score [the test score] shows what their true ability is." This is a rare example of Code 4, relating an interval to estimation theory and the distance between an estimate and the true parameter. In general, this student was not clear whether it was the population or sample mean that was falling into the interval. What was consistent was that a fixed interval was drawn or that a fixed interval existed. The strongest evidence for this was that in calculating an interval, the student began by drawing a normal distribution and demarking the acceptance region on the figure—clear examples of Codes 7 and 8 (Fixed Disk). As this interviewee wavered between population and sample means, it was difficult to determine which, if either, was fixed in her mind. This is an interesting situation that could lead to additional coding in future studies. Without further explanation, the student's response would indicate either that the population mean was moving, which is an example of Code 6, or that it was an example of a fixed sample mean, which does not currently have a code. This student made more references to the Fixed Disk metaphor but was not definitive in selecting a graphical representation in Question 3. Instead of selecting the representation with μ falling into a fixed interval as one might predict, the student explained both representations and the circumstances in which one would select one or the other. At this point, she correctly captured the essence of both metaphors, and only when forced to answer did she pick the correct representation.

The third interviewee provided a contrasting set of results as this student appeared to rely on imagery consistent with the Changing Ring metaphor. In explaining a confidence interval, she generally avoided language like "the mean falls in the interval," instead using phrases like "this mean lies between these two values." The shift to a more passive verb—*lies* in place of *falls*—may indicate that the student was imagining a stationary population parameter. That this language was clearly her creation is also important. Of great potential to further studies is that her reasoning was to contrast confidence intervals with acceptance regions. She even pointed out that her confusion was caused by the juxtaposition of acceptance regions and confidence intervals in instruction and texts, stating, "I think there are two things that you figure in a row, and the second one is the confidence interval and the first one is the rejection region or something." While coded as support for the Changing Ring metaphor (Code 8) in the current study, this remark raises the question of whether there is a need for two separate codes—one for when a student contrasts confidence intervals with hypothesis testing and one for when the confuses the two methods.

The third student's response—contrasting confidence intervals with acceptance regions—raises two additional points of interest. First, as confidence intervals are replacing *p*-values as the preferred method of reporting experimental results, they are becoming a tool for inference, while their historical origin is in estimation and bounding the distance between an estimate and the true parameter (Stigler, 1986). Second, the student's response shows how the close sequencing of the topics of confidence intervals and rejection regions in contemporary statistics textbooks (e.g., Wardrop, 1995; Marascuilo & Serlin, 1988) blurs the distinctions between them for students at this stage of learning.

Conclusions

The study reported here documents the existence of at least two conceptual metaphors for confidence intervals, one consistent with the view held by the statistics community and one at odds with that view. The Changing Ring metaphor is consistent with the generally accepted reasoning in statistics since the interval is determined from the sample and changes from sample to sample, and the fixed point is the population parameter, which, while often unknown, is invariant. This is not the case for the Fixed Disk metaphor, in which the disk representing the confidence interval does not change, and the population mean, which is fixed, takes on a more dynamic characteristic as it moves from sample to sample. Our results show that gestures and speech are mechanisms that can illuminate students' understanding of confidence intervals—and in particular, students' tendency to move between these two incompatible conceptual metaphors. One student in our study seemed to follow the logic of the Fixed Disk metaphor. After being prompted to push the metaphor further, she was forced to say something she knew was incorrect. Then she appealed to the Changing Ring metaphor, but was uncomfortable with it and subsequently returned to the Fixed Disk metaphor. The second student began with the Fixed Disk metaphor and switched to the Changing Ring metaphor as she became more comfortable reasoning through the questions. She ultimately was unsure of her answers, but her reasoning referred to the Changing Ring metaphor. The final student referred to the Fixed Disk metaphor only as a source of contrast, bringing to light a limitation of our scoring system—namely, that it treats references to the Fixed Disk metaphor the same whether they result from confusion and misapplication or serve as a contrast with the correct interpretation. In general, this work demonstrates that, while different students use these two competing metaphors in different ways, the two metaphors are present in their reasoning.

Educational and Scientific Importance

This study provides information to assist in the instruction and assessment of confidence intervals for graduate students in the behavioral and social sciences and to improve the reporting of experimental results in these disciplines. As noted, contemporary statistics textbooks often present the topics of confidence intervals and hypothesis testing in close sequence. While linking these topics is convenient due to their computational similarities, it comes at the cost of conceptual understanding: there is evidence that the correct metaphor for confidence intervals, while present in students, is not dominant. Furthermore, we found evidence that the typical curriculum sequence may be causing an even more fundamental problem by confusing sample and population characteristics. Our study demonstrates that confidence intervals are not interpreted consistently by budding researchers in the education sciences. As there is growing pressure to report research findings with confidence intervals, this may lead to considerable miscommunication as social scientists read and contribute to the research base. Instruction that better distinguishes between confidence intervals and hypothesis testing might make students more comfortable with the difference between these two superficially related topics.

This work provides a starting place for discussions and studies on how students are conceptualizing statistical reasoning. Most important is the demonstration that gestures and speech can reveal aspects of the conceptual models that students employ during statistical reasoning. Having established some of the characteristics of these conceptual metaphors, this study lays the basis for future research using surveys as a method for studying a significantly larger population of students to determine how widespread the two metaphors are. Furthermore, this study shows how the coding system could be improved in future research to recognize the distinction between confusing metaphors and contrasting metaphors. Future study might also identify additional items that should be a part of the coding system. At this point, this study has demonstrated that students' conceptual metaphors can be identified through speech and gesture and that multiple metaphors are competing in some students' minds when they reason about confidence intervals.

Garfield and Ben-Zvi (in press) also presented a research-based approach to statistics education that elicited students' intuitions about variability and inference early in the course and then revisited them throughout in order to provide a strong conceptual bridge to formal ideas as they were developed later in project- and simulation-based activities. For example, before introducing the formulas for determining the intervals, the authors presented visual displays of computer-generated confidence intervals to show students how they could change from sample to sample. This approach is consistent with the findings of this study and the conceptual metaphors presented, particularly in reference to the fact that confidence intervals are sample statistics.

Many extensions to this work are worth pursuing. The next immediate step is to design a survey instrument that might identify which metaphor is dominant in a student's thinking. Additionally, identifying correlations between dominant metaphors and the known fallacies about confidence intervals could further identify specific points to be addressed and indicate how common confidence intervals are misunderstood.

References

Alibali, M. W., Bassok, M., Solomon, K., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science, 10,* 327–333.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5^{th} ed.). Washington, DC: American Psychological Association.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396.

delMas, R., Garfield, J., & Chance, B. (in press). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*.

Franklin, C., & Garfield, J. (2006). The GAISE project: Developing statistics education guidelines for preK–12 and college courses. In G. F. Burrill, & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance: 2006 NCTM yearbook* (pp. 345–375). Reston, VA: National Council of Teachers of Mathematics. Retrieved June 5, 2008, from http://www.amstat.org/Education/gaise/GAISECollege.htm

Garfield, J., & Ben-Zvi, D. (in press). *Developing students' statistical reasoning: Connecting research and teaching practice*. Emeryville, CA: Key College.

Garfield, J., delMas, B., & Chance, B. (n.d.). *Some common misconceptions students should NOT have*. Retrieved June 5, 2008, from Tools for Teaching and Assessing Statistical Inference Web site: http://www.tc.umn.edu/~delma001/stat_tools/

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20,* 1–55.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think.* Cambridge, MA: Harvard University Press.

Lakoff, G., & Nunez, R. E. (2000). *Where mathematics comes from*. New York: Basic Books.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences.* New York: W. H. Freeman.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Mills, J. D. (2002, March). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1). Retrieved June 5, 2008, from http://www.amstat.org/publications/jse/v10n1/mills.html

Nathan, M. J. (in press). An embodied cognition perspective on symbols, grounding, and instructional gesture. In M. DeVega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment and meaning: A debate*. Oxford, UK: Oxford University Press.

Nathan, M. J., & Bieda, K. N. (2006). *What gesture and speech reveal about students' interpretations of Cartesian graphs: Perceptions can bound thinking* (WCER Working Paper Series #2006-2). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved June 5, 2008, from http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Rossman, A., & Chance, B. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *College Mathematics Journal, 30*(4), 297–305. Retrieved March 17, 2007, from http://rossmanchance.com/papers/topten.html

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55,* 182–186.

Stigler, S. H. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Belknap Press of Harvard University Press.

Thompson, P. W., Saldanha, L. A., & Liu, Y. (2004, April). *Why statistical inference is hard to understand.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wardrop, R. L. (1995). *Statistics: Learning in the presence of variation*. Dubuque, IA: William C. Brown.

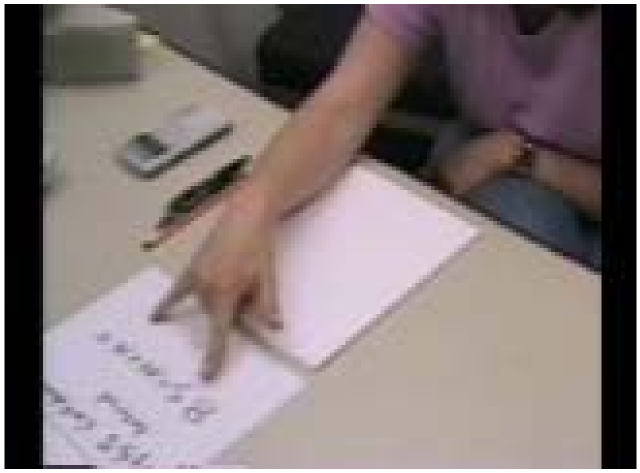


Figure 1. **Gesture showing the fixed boundary points of a confidence interval in Question 1 (see Appendix), as evidence of the incorrect conceptual metaphor.**

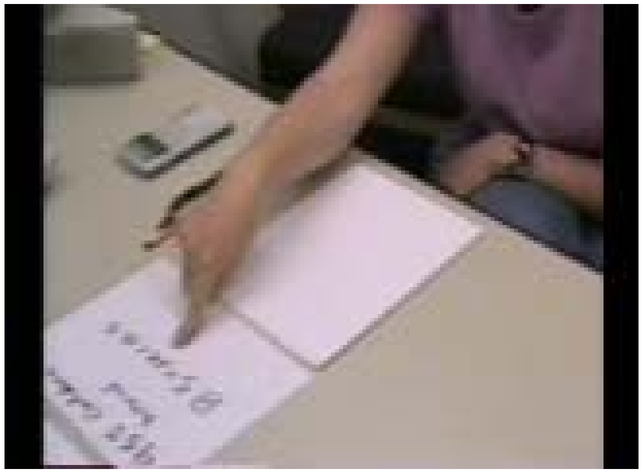


Figure 2. **Gesture indicating a fixed population parameter by pointing and anchoring the mean, generally evidence of the correct conceptual metaphor.**

Figure 3. **Student using an open-hands gesture to indicate a continuum or many possible points. Gesture coded as changing ring metaphor.**

Figure 4: **Student using hands to mark fixed boundaries or a specific interval. Width of gesture can be used as an indication of confidence. Gesture coded as fixed disk metaphor.**

Appendix

Interview Items

1) Describe to a student new to statistics what a confidence interval is. Explain the following confidence interval:

95 % confidence interval

$$8.5 < \mu < 11.5$$

2) A researcher is studying the amount of selenium people are getting in their diets. He surveys people at a student union, and based on the food they have had in the last 24 hours, he collects the following data:

$$\overline{X} = 189$$
$$\sigma = 42$$
$$n = 36$$
$$Z_{\alpha/2} = 2$$

Construct a confidence interval. To assist you, here is the formula:

$$\overline{X} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

ANSWER:

$$175 < \mu < 203$$

A second researcher conducts a similar study on the same population and gets the following data:

$$\overline{X} = 191$$
$$\sigma = 36$$
$$n = 81$$
$$Z_{\alpha/2} = 2$$

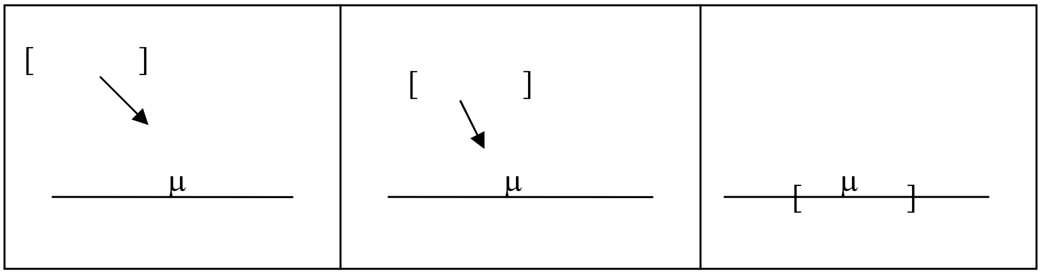
Construct a second confidence interval.

ANSWER:

$$183 < \mu < 199$$

3) Which of these two graphical depictions reminds you more of a confidence interval?

a)



b)