Continued misinterpretation of confidence intervals: response to Miller and Ulrich

Richard D. Morey
Cardiff University

Rink Hoekstra
University of Groningen

Jeffrey N. Rouder
University of Missouri

Eric-Jan Wagenmakers
University of Amsterdam

Miller and Ulrich (in press) critique our claim (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), based on a survey given to researchers and students, that even seasoned researchers have difficulty interpreting confidence intervals (CIs). They suggest that survey respondents may have interpreted the statements in the survey that we deemed incorrect in an idiosyncratic, but correct, way, thus calling into question the conclusion that the results indicate that respondents could not properly interpret CIs. We show, however, that their "interpretations" are trivial restatements of the definitions of confidence procedures; further, although they are tautologically correct by themselves, they cannot be correct interpretations of the statements in the survey. Miller and Ulrich have unintentionally underscored our main point: confidence intervals are counter-intuitive, and even seasoned researchers get it wrong.

DRAFT

Knowing how to interpret confidence intervals is critical to researchers who encounter and use confidence intervals, and in debates over the usefulness of various statistical methods. In our paper (Hoekstra et al., 2014), we showed that incorrect interpretations of confidence intervals (CIs) were frequently endorsed by researchers, master students, and by first year students with limited knowledge of inferential statistics. We concluded that these errors in interpretation imply a widespread misunderstanding of the role, affordances, and use of CIs. Miller and Ulrich (in press), however, argue that some of the statements we considered incorrect could in fact be considered correct, given a particular interpretation of probability and confidence. We argue their argument is flawed, and that Miller and Ulrich (henceforth, MU) unintentionally underscored our case that the interpretation of CIs is conceptually difficult and often misunderstood.

MU's argument fails for three reasons: first, the interpretation of confidence intervals they suggest is actually no interpretation at all, but instead is a trivial restatement of the definition of a confidence interval; second, to adopt their interpretation as justifying the probability statements we labeled as incorrect would lead to contradiction of the laws of probability and non-unique assignments of confidence (regardless of the interpretations of those terms); and third, that their faith that writers on statistics truly understand confidence intervals is misplaced.

We begin by drawing a helpful distinction between a confidence procedure (CP) and a confidence interval; a confidence procedure generates confidence intervals when supplied with data; a 95% confidence procedure is a procedure that generates intervals that contain the true value of interest 95% of the time. A confidence interval is a realization from a confidence procedure. The basic defining property of confidence intervals as realizations of procedures was introduced by Neyman (1934, 1937) who laid the foundations of confidence interval theory. MU do not disagree with us on the definition of confidence intervals; their disagreement with us involves what interpretations confidence intervals are acceptable. It is worth emphasizing that the points we make about confidence intervals, and probability in general, are not controversial from the the point of view of statistical theory; they have been emphasized by frequentist and Bayesian philosophers and statisticians since the inception of CI theory in the 1930s. Readers interested in a more thorough treatment of CI theory are referred to Morey, Hoekstra, Rouder, Lee, and Wagenmakers (unpublished manuscript).

A trivial interpretation

MU take issue in particular with several statements in our survey that we said were incorrect; they assert that if the survey respondents had a definition of probability other than a frequentist one, our statements 1, 3, 4, and 5 need not be read as incorrect. MU offer alternative statements 1', 3', and 4' that they say could be the way that respondents interpreted

Address correspondence to Richard D. Morey (richarddmorey@gmail.com). We thank All code and the source for this document are available at https://github.com/richarddmorey/MillerUlrichResponse. Draft date: July 1, 2015.

confidence intervals (pps 7 and 14). MU suggest that perhaps survey respondents understood, for example, 4 as 4' and, because 4' is not incorrect, this should not be taken as evidence that they misunderstand confidence intervals. MU's alternative statements have the interesting property that they are all trivially true, regardless of the observed data: in fact, they are mere restatements of the definition of a confidence procedure. Although MU state that "Any interpretation of sample data should in some way summarize the information provided by the sample" (p. 6), their correct "interpretations" amount to tautologies that do not summarize the information provided by the data in any way.

Consider MU's statement 4' for a $CI_{95\%}$ of [.1, .4] ("If the current sample is one of the 95% of all samples with relatively small values of $|\bar{X} - \mu|/S$, then μ lies in the interval 0.1−0.4", p. 7). Let "relatively small" mean that $|\bar{X}-\mu|/S < c$ for some positive value c. Then a bit of algebra yields

$$\bar{X} - cS < \mu < \bar{X} + cS.$$

But c must be chosen such that 95% confidence procedure results. This condition, then, is merely the condition that μ is in the interval. MU's statement 4' amounts to "If μ is in the interval, then μ is in the interval, and μ will be in the interval 95% of the time." The first part of the statement is an information-free tautology; the only non-trivial information we are left with is "μ will be in the interval 95% of the time." Of course, this is merely the definition of the confidence procedure, and makes reference only to the long-run property of the interval. Contrary to MU, statements 1' and 3' and 4' are not "entirely appropriate conclusions"; they are, rather, entirely unhelpful tautologies.

We find it implausible that respondents would interpret the statements as MU suggest, for three reasons. First, a MU's interpretations would have strange implications for other statements respondents could also endorse. Consider the situation where one has perfect knowledge of μ (say, through simulation, theoretically, or with sufficient previous data, such as modern measurements of the speed of light in a vacuum). This would not change MU's probability statements at all; even though we might know for certain that, say,

- (Statement A) $\mu = 2$,

and upon observing a $CI_{95\%}$ of [.1, .4], MU would also tell us we could state that

- (Statement B) $Pr(.1 < \mu < .4) = .95$

These two statements would be contradictory under any formal definition of probability, including Bayesian and frequentist ones; to see why, replace μ with the known value of μ. If one accepts MU's argument, however, there is no contradiction in endorsing both A and B simultaneously[1]. That 4' is true regardless of the true value of μ — whether it is known or not — and true for any observed CI makes 4' is a trivial, useless statement.

If MU are correct that people interpret CIs this way, then a modified version of the survey could be circulated in which in information is added that $\mu = 2$, with certainty, and this should produce no change in the responses among those who interpret 4 as 4'. We suspect, however, that this would have drastic effects on the responses, because most people would find the two statements A and B above to be contradictory. Anyone who applied the probability knowledge from their introductory course in statistics would find the statements contradictory. If Statements A and B appear to a respondent as contradictory, they cannot be interpreting the statements as MU suggest.

Second, we note that even Dempster (1964) — cited by MU as clarifying the interpretation they suggest — writes that "*[i]t does not appear to be widely understood that, after the observation is taken, the defining property [of the confidence interval] admits only a postdictive interpretation [that is, the trivial one suggested by MU].*" (p. 57; emphasis in original). Dempster, in fact, emphasizes this sentence for effect. He is talking about mathematical statisticians, and explicitly saying that they do not appear to understand confidence intervals; in particular, that they only admit trivial statements (this is not cited as a point of strength for CIs by Dempster). Contrast this with MU's statement that "it is implausible that such well-established mathematical statisticians do not understand CIs." Dempster apparently thought that these misunderstandings were widespread and important enough to emphasize. In fact, the issues with confidence interval interpretation are well-known among writers in statistics. Neyman (1952, pp. 211-215) presents a humorous, fictional dialogue in which an "eminent elderly statistician" has great difficulty understanding confidence intervals. Mayo (1981), critiquing a paper by Seidenfeld — one of the leading figures in the philosophy of probability — writes that "while confidence levels are often (wrongly) interpreted as providing [...] a measure of [certainty that the parameter is in the interval], no such interpretation is warranted. Admittedly, such a misinterpretation is encouraged by the word 'confidence'." Neyman, Dempster, and Mayo do not take it for granted that statisticians understand confidence intervals; we see no reason why MU should take it for granted either.

A third reason we believe that respondents were unlikely to be adopting MU's trivial interpretation is that they do not appear to have any implication for where one should believe the parameter actually is. Why would one adopt CIs if such trivial interpretations were the only possible ones? Dempster

[1]This may seem counter-intuitive, but the only way that a logical conditional statement such as MU's 4' can be false is if the antecedent is true and the consequent is false. If $\mu = 2$ is true and the CI is [.1, .4], then the antecedent and consequent in statement 4' are both false, and hence 4' is still true.

points out that researchers actually desire a *different* sort of probability statement than the one MU suggest: a so-called "predictive" probability statement, one that has implications for where the parameter is believed to be. He says "I find Neyman quite vague on the intellectual mechanism whereby [CI] interpretations come to have predictive implications. I suspect that particular observed confidence statements are intended to have the effect of predictive probability statements, without actually using the word probability and without paying the price which the use of the word probability demands." (p. 60).[2] We completely agree; the advocacy of confidence intervals rests on an interpretive sleight of hand by which trivial statements are rendered as predictive statements about the parameter (see also Morey et al. (unpublished manuscript)).

Finally, while we do not believe that it likely that respondents were interpreting "probability" as MU suggest, we agree with MU that participants may have a different understanding of probability than the frequentist one that serves as the basis of CI theory. We believe it is more likely that respondents interpreted the word "probability" as a degree of belief. All of MU's examples (the coin flip, the deck of cards, and the random-number generator) are consistent with this interpretation of probability. This is not helpful to MU's case: even if the respondents did adopt this definition of probability, it does not affect our conclusions about the incorrectness of the statements in our questionnaire because there is no necessary relationship between the CI and the stated degree of belief.

Wrong under any definition of probability

Statements 1, 3, and 4 are incorrect by any commonly taught definition of probability, and MU can only defend these statements by asserting that respondents may be understanding these statements as if they are tautological restatements of the definition of a confidence procedure, and hence not wrong. There are good reasons to believe that their trivial interpretations are not widespread, for reasons including those listed above. At any rate, MU offer no empirical evidence that their trivial interpretation is widespread; in fact, as MU themselves point out, nonsensical statement 6 was endorsed at a similar rate as statement 4, suggesting that respondents do not understand CIs at all. Furthermore, respondents who endorsed 4 were barely less likely to endorse statement 6 than those that did not endorse statement 4 (62% vs 65%), suggesting that endorsement of statement 4 does not reflect any substantial knowledge of confidence intervals.

Suppose, however, that MU are correct, and respondents were substituting tautologically true statement 4' for our statement 4. Is 4' a suitable substitution for 4? The answer is "no", regardless of one's interpretation of the word probability *so long as that definition conforms to the requirements of the laws of probability.* To see why, one must understand

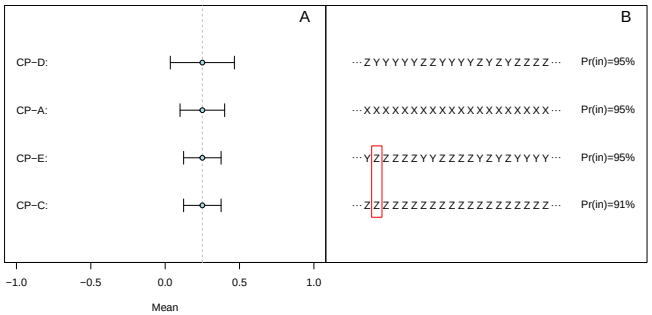

Figure 1. The reference class problem. A: Three nested confidence intervals, all computed from confidence procedures with the same 95% probability of including the true value. B: Excerpts from four sequences of applications of the confidence procedures in Panel A. Each letter represents computation of an interval with a different confidence coefficient; "X" represents a 95% interval, "Y" represents a 99% interval, and "Z" represents a 91% interval. The rectangle shows how the same confidence interval can be seen as a sample from two sequences, each with different probabilities of containing the true value.

two facts: first, there are many confidence procedures for the same problem, all with the same confidence coefficient but yielding different intervals; second, a given confidence interval can arise from two different confidence procedures, each with different confidence coefficients. This lack of unique inferences has been known since the inception of confidence interval theory. Fisher, for instance, in the discussion of the Neyman's first paper including CI theory, stated that CI theory "had been erected at considerable expense, and it was perhaps as well to count the cost. The first item to which [Fisher] would call attention was the loss of uniqueness in the result, and the consequent danger of apparently contradictory inference." (Neyman, 1934, discussion at p. 618).

We can demonstrate this using multiple confidence procedures. Suppose we have a method for constructing CIs with any given confidence coefficient: our only requirement is that the width of the CI is an increasing function of the confidence coefficient, which is true of most confidence procedures in common use. We denote as Confidence Procedure A (CP-A) a CP constructed with a confidence coefficient of 95%. We can construct another 95% confidence procedure by randomly combining two confidence intervals with different confidence coefficients. For instance, let CP-B and CP-C be intervals with $\alpha = .01$ and .09, respectively; that is, CPB and CPC are respectively 99% and 91% confidence procedures. We define two more 95% confidence procedures, CP-D and CP-E:

[2] Compare this to MU's statement that "the word 'probability' is conspicuously absent from [our] interpretations."

- CP-D: Flip a fair coin. If the coin shows "heads" then compute an interval from CP-B; else, compute an interval from CP-C.

- CP-E: Flip a fair coin. If the coin shows "tails" then compute an interval from CP-B; else, compute an interval from CP-C.

Procedures CP-D and CP-E are easily seen to be 95% confidence procedures. They may not be good confidence procedures; however, this is not at issue here.[3] Our analyst also needn't know how these CIs were constructed; indeed, this information was missing, purposefully, from the survey. Rejecting CIs on either of these grounds would be a tacit admission that some other information is needed to interpret a confidence interval besides knowledge that an interval was computed from some confidence procedure, which is precisely our point; more importantly, this point has been made by decades of statistical literature on CIs.

Notice that one of CP-D or CP-E will always yield the wider interval of CP-A, CP-D, or CP-E, and the other the shorter; note also that CP-A will always be nested within the wider of CP-D and CP-E, and the shorter will always be nested within CP-A. This situation is depicted in Figure 1A. Neyman (1952) used similarly nested confidence intervals to warn about just the misinterpretations that MU make.[4] If MU are correct that merely knowing that an interval was computed from a confidence procedure with 95% coverage licenses the statement that "the probability that the interval contains the true value is 95%", then the law of total probability implies that if any one of the nested 95% intervals contains the true value, then the shortest of the 95% intervals must contain the true value. Because the laws of probability will be the same for any interpretation of probability, no reinterpretation of the survey questions can avoid it. Moreover, a 95% confidence procedure can always be constructed that sometimes yields arbitrarily large or small intervals, which would imply that the interval must be in the arbitrarily small interval. This is, of course, absurd.

But there are even bigger problems: if MU are correct, any given confidence interval can be assigned multiple probabilities simultaneously (Figure 1B). To see why, consider the trio of confidence procedures CP-B, CP-D, and CP-E. Half of the time, CP-C and CP-E will yield the same interval. But CP-C is a 91% procedure, and CP-E is a 95% confidence procedure. Under MU's logic, we can assign the same interval both 95% and 91% probability. Similarly, as previously noted, with nested CIs, if MU are correct then the laws of probability imply that all 95% probability is located in the shortest interval of CP-E or CP-D, and hence the law of total probability implies that we should be able to assign 95% probability to the shorter of the two intervals. However, always taking the shorter of these two is obviously a 91% confidence procedure. If MU were correct, we can could assign both 95% and 91% probability to the shorter interval from CP-D and CP-E. The laws of probability, however, require unique probability assignments; therefore, MU's proposed assignment of probability to individual intervals lapses into contradiction, *regardless of the interpretation of probability one chooses.*

These points are not controversial among statisticians and philosophers of statistics; these are precisely the sorts of contradictory inferences that Fisher warned about. In fact, the problem of non-unique probability assignments is an example of a more general and well-known problem called the "reference class problem" (Venn, 1888; Reichenbach, 1949): any CI can be thought of as being a sample from multiple different long-run sequences, and hence any probability assignment to individual CIs will not be unique. This is perhaps why Neyman (1952) specified that, "[all the CI] does assert is that the probability of success in estimation using either of [...] two formulae is equal to [95%]" (p. 214), avoiding the reference class problem and non-unique probability assignments because probability is never attached to any particular interval – only procedures.

Since MU's defenses of statements 1, 3, and 4 ultimately fail, we are left with the last statement they defended: statement 5 ("We can be 95% confident that the true mean lies between 0.1 and 0.4"). The correctness of this statement will depend, in some sense, on the definition of "confident." There are two immediately obvious definitions of the word "confidence" that one can use to interpret the statement. The first is the colloquial meaning, in which "confidence" entails a degree of belief. In the context of CIs, this interpretation of confidence would require that there is a 95% belief that the parameter is in the interval $[0.1, 0.4]$. Within frequentism, there is no formal justification for quantifying belief, and without any information about the context, no reasonable prior that would justify a Bayesian interpretation can be constructed. In fact, it is possible to construct confidence procedures — even good frequentist procedures — for which it is sometimes known with certainty that a computed interval does, or does not, contain the true value (Berger & Wolpert, 1988; Morey et al., unpublished manuscript). If confidence and degree of belief were synonymous, we would simultaneously assign 95% and either 0% or 100% certainty to such an interval. It is hard to imagine any reasonable way of assigning degrees of belief that would allow this. The colloquial interpretation of CIs is clearly not a valid justification for statement 5.

MU, however, present a second interpretation of "confi-

[3] As MU point out (footnote 1, p. 5), the relative performance of various CIs is not relevant for the present discussion; all that is relevant is that a procedure is a confidence procedure.

[4] Examples of nested confidence intervals are easy to construct. For examples, see Neyman (1952), Welch (1939), and Morey et al. (unpublished manuscript).

dence", which is the technical, frequentist definition of "confidence." We argued that this can only refer to the procedure and not to the specific interval. MU assert that "by this argument, a correct interpretation of a CI can never mention the numerical bounds of the CI that are computed from the sample. But such an argument is clearly inconsistent with [Neyman's] stated purpose of CIs..." (p. 12). MU's assertion here is not true; the bounds can be mentioned, but only as part of the *decision* entailed by the procedure to act as though the parameter is in the interval, *not* as an interpretation in terms confidence or probability. MU's assertion is also incorrect as an interpretation of Neyman's purpose for CIs. Notice that Neyman (1937) does not say that the analyst should be 95% "confident" that the interval contains the parameter; he says that the analyst should state (and act as though) the parameter is *definitely in the interval.* Neyman (1941) makes this very clear:

> "[I]t is not suggested that we can 'conclude' that [the interval contains μ], nor that we should 'believe' that [the interval contains μ]...[we] decide to behave as if we actually knew that the true value [is in the interval]. This is done as a result of our decision and has nothing to do with 'reasoning' or 'conclusion'. The reasoning ended when the [CI procedure was derived]. The above process [of using CIs] is also devoid of any 'belief' concerning the value [...] of [μ]." (p. 133-134)

The level of confidence is an assertion about the procedure, as Neyman repeatedly stated. If there were any definitions of "confidence" that would render statement 5 acceptable, they would have to be consistent with non-unique assignments of confidence, due to the reference class problem. Unlike with "probability," however, there is no rule saying that we cannot have non-unique "confidence," if we allow confidence to mean anything we like. This may save statement 5, but it also means that any other similar statements with "confidence" other than 95% would also be simultaneously implied. So statement 5 is either not true, or trivially true because any such statement would be true. To the extent that respondents would not also have endorsed the statement that "We can be 1% confident that the true mean lies between 0.1 and 0.4," they lack a justification for statement 5. To the extent that they would have simultaneously endorsed all such statements, the assessment of confidence carries no information and is hence meaningless. We, like MU, suspect that an endorsement of statement 5 was not actually an endorsement of the claim that CIs are meaningless, and so it must lack justification from the information given and hence be incorrect.

Misplaced faith

MU cite others who present statement 5 as the interpretation of a CI as evidence for their case that this interpretation must be correct. They consider it "implausible that such well-established mathematical statisticians do not understand CIs". MU are right that many papers and textbooks show interpretations of CIs which we would consider incorrect, and some are indeed written by well-known statisticians. Given the previous demonstrations that p values are often misinterpreted (e.g., Haller & Krauss, 2002), it should come as no surprise that CIs are also misinterpreted. Moreover, as previously mentioned, theoreticians such as Dempster, Neyman, and Mayo seemed to be aware that even mathematical statisticians could have trouble interpreting confidence intervals, so we do not understand why MU find this "implausible". But if the interpretations are demonstrably problematic — and they are, as we show and as even a cursory reading of the theoretical CI literature will affirm — then this argument carries no weight. It amounts to a mistaken argument from authority.

There are many plausible explanations why textbook authors would include incorrect statements about CIs. The most obvious one is that CIs are indeed very hard to grasp, even for statisticians, which is underscored by this very discussion. Some authors explicitly struggle with the discrepancy between use of CIs in practice and the theory behind it. **Howell:2012** for example, states that "So what it does it mean to say that the 95% confidence interval is $1,219 \leq \mu \leq 1,707$? For seven editions of each of two books I have worried and fussed about this question" (p. 194). Moreover, textbooks authors may "simplify" the interpretation for didactic reasons, because they think the correct interpretation is too difficult to understand for students. Readers, after all, often do not appreciate difficult explanations. It is reasonable to think that textbooks suffer from the same problem as science journalism (D. F. Ransohoff & R. M. Ransohoff, 2001): textbook authors are rewarded for the apparent clarity that often arises from oversimplification, not necessarily for correctness. Finally, textbook authors may try to connect to what is typically assumed in practice, rather than what is theoretically correct. An attention to correctness is often assumed to indicate a lack of "pragmatism" when what is widely believed is different from what is correct.

In summary, the alternative interpretations MU presented are unlikely, and, more importantly, that they are incorrect independent of the definition of probability. For these reasons, all six statements in our questionnaire are false, and the main conclusion that apparently researchers find it hard to interpret CIs remains valid. This conclusion is underscored by the fact that the two items that MU would consider false as well were endorsed in similar proportions as those items who MU thought were correct, and by the fact that MU defended several of these false statements. Our finding may be

shocking given the heavy endorsement of CIs in the last few decades, but apparently even methodologists have difficulty agreeing that all the items were presented were false. Finding consensus among researchers might probably have been an even more surprising finding.

References

Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)* Hayward, CA: Institute of Mathematical Statistics.

Dempster, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *Journal of the American Statistical Association*, *59*(305), pages. Retrieved from http://www.jstor.org/stable/2282858

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online*, *7*.

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

Mayo, D. G. (1981). In defense of the Neyman-Pearson theory of confidence intervals. *Philosophy of Science*, *48*(2), 269–280.

Miller, J. & Ulrich, R. (in press). Interpreting confidence intervals: a comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review*.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (unpublished manuscript). *The fallacy of placing confidence in confidence intervals*. submitted. Retrieved from http://dx.doi.org/10.5281/zenodo.16991

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625. Retrieved from http://www.jstor.org/stable/2342192

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*, 333–380. Retrieved from http://www.jstor.org/stable/91337

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, *32*(2), 128–150. Retrieved from http://www.jstor.org/stable/2332207

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. Washington, D.C.: Graduate School, U.S. Department of Agriculture.

Ransohoff, D. F. & Ransohoff, R. M. (2001). Sensationalism in the media: when scientists and journalists may be complicit collaborators. *Effective Clinical Practice*, *4*, 185–188.

Reichenbach, H. (1949). *The theory of probability*. Berkeley, University of California Press.

Venn, J. (1888). *The logic of chance* (third edition). London: Macmillan. Retrieved from https://archive.org/details/logicofchance029416mbp

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *The Annals of Mathematical Statistics*, *10*(1), 58–69. Retrieved from http://www.jstor.org/stable/2235987