QUALITATIVE RESEARCH: AN ESSENTIAL PART OF STATISTICAL COGNITION RESEARCH

PAV KALINOWSKI
Statistical Cognition Laboratory, School of Psychological Science, La Trobe University
p.kalinowski@latrobe.edu.au

JERRY LAI
Statistical Cognition Laboratory, School of Psychological Science, La Trobe University
kj2lai@students.latrobe.edu.au

FIONA FIDLER
Statistical Cognition Laboratory, School of Psychological Science, La Trobe University
f.fidler@latrobe.edu.au

GEOFF CUMMING
Statistical Cognition Laboratory, School of Psychological Science, La Trobe University
g.cumming@latrobe.edu.au

ABSTRACT

Our research in statistical cognition uses both qualitative and quantitative methods. A mixed method approach makes our research more comprehensive, and provides us with new directions, unexpected insights, and alternative explanations for previously established concepts. In this paper, we review four statistical cognition studies that used mixed methods and explain the contributions of both the quantitative and qualitative components. The four studies investigated concern statistical reporting practices in medical journals, an intervention aimed at improving psychologists' interpretations of statistical tests, the extent to which interpretations improve when results are presented with confidence intervals (CIs) rather than p-values, and graduate students' misconceptions about CIs. Finally, we discuss the concept of scientific rigour and outline guidelines for maintaining rigour that should apply equally to qualitative and quantitative research.

Keywords: *Statistics education research; Mixed methods; Scientific rigour; Qualitative analysis*

1. MIXED METHODS IN STATISTICAL COGNITION

Statistical cognition refers to "the cognitive processes, representations, and activities involved in acquiring and using statistical knowledge," as well as the research program that investigates these processes (Beyth-Marom, Fidler, & Cumming, 2008, p. 22). In this way statistical cognition is similar to the discipline of *cognition*, which refers to both mental processes and the body of research investigating these processes. In this paper we describe how both quantitative and qualitative methods are used together in our statistical cognition research program.

Regardless of whether research is quantitative or qualitative, we believe that researchers should describe the context of their work and their preconceptions and assumptions. For this reason, we begin this paper by stating that we are advocates of statistical reform in psychology; that is, we believe that the dichotomous thinking associated with Null Hypothesis Significance Testing (NHST) has damaged the progress of psychology and that estimation-based techniques, that is, effect sizes and confidence intervals (CIs), are better tools for statistical communication. However, we also believe that statistical reform should be evidence-based. As such, we believe that advocates of reform should provide empirical evidence that the alternatives to NHST that they promote are better communicators of inferential information and less prone to misinterpretation and misuse. Our statistical cognition program has produced some evidence in favour of CIs (e.g., Fidler & Loftus, 2009), but the four studies recounted here show that collecting such evidence is by no means straightforward!

Qualitative research is essential in fulfilling the goals of the statistical cognition program in at least two ways. First, it helps achieve fuller and more complete descriptions of phenomena. We illustrate this in the first two of our four studies: Fidler, Thomason, Cumming, Finch, and Leeman (2004) used a mixed approach to examine the effect of error bars in result interpretation in medical journals. Faulkner (2005) used interviews to explore students' preference and efficiency in interpreting CIs and NHST.

Secondly, qualitative methods may be very useful in suggesting new directions for research. Our exploratory studies, open-ended questions, and interviews have yielded unexpected and novel insights and have led to new research programs. Again, two studies are offered as examples: Coulson, Healy, Fidler, and Cumming (2010) produced unexpected results when comparing researchers' interpretations of NHST and CIs, which led to a new research program. Kalinowski (2010) explored student misconceptions of CIs using both qualitative and quantitative methods.

Of course, qualitative methods have more to offer than just these two features (more complete description and new directions). In our account of the four studies that follows we will also illustrate how qualitative methods have helped correct our misinterpretations of quantitative results, and in other cases provided triangulation. Statistical reasoning is often fragile, and quantitative methods can fail to capture subtleties and layered misconceptions. For example, a quantitative survey may provide an indication of how many students have a false belief about some statistical concept, but not necessarily how they arrived at that false belief, or which other statistical concepts might be implicated. Qualitative methods can help us access processes and the mental models at work in the formation of misconceptions.

Finally, we will address the issue of robustness in qualitative research. Qualitative methods are often mis-associated with terms such as *subjective* or *biased*. In reality, research judgment is an integral and important part of both quantitative and qualitative methods. In the final section of this paper we will explicate established guidelines (namely those of Elliott, Fisher, & Rennie, 1999) for maintaining rigour in qualitative research and argue that the same standards should also be expected of quantitative research.

2. ACHIEVING MORE COMPLETE DESCRIPTIONS OF PHENOMENA: FIDLER ET AL. (2004)

As mentioned above, one major goal of statistical reform in psychology is the replacement of NHST p-values with CIs. A common way to examine reform progress is via journal surveys on the prevalence of reporting practices (e.g., Cumming et al., 2007;

Thompson & Snyder, 1997). Such surveys provide quantitative estimates of the extent or lack of change in statistical practice.

In psychology, such journal surveys have consistently demonstrated little change in response to reformers' calls for downplaying NHST. In medicine, by contrast, changes have been reasonably dramatic, starting in the mid-1980s when several journal editors enforced new reporting policies. Fidler et al. (2004) investigated changes in medicine by surveying statistical practices in two medical journals, the *American Journal of Public Health* (*AJPH*) and *Epidemiology*. Both journals were subject to strict editorial policies from then-editor Kenneth Rothman that eschewed p-values and encouraged use of CIs.

Quantitative The quantitative component of this study recorded the proportion of articles reporting p-values versus CIs. Results revealed a dramatic increase in the uptake of CIs under Rothman's editorship—from 10% pre-Rothman (1983) to 60% at the peak of his influence (1987). There was a corresponding drop in p-value reporting: from 63% in 1982 to just 6% in 1986–1989. In *Epidemiology*, the influence of Rothman's policy was even more striking: 94% of articles reported CIs in 2000 and none reported p-values. From the quantitative survey alone it seemed that statistical reform in medicine had been quite successful.

Qualitative The qualitative component examined the interpretation of results, in particular, how the increase in CI reports changed the way authors discussed their results. Did they now reflect on the width of the CI and talk about issues of statistical power/precision (we know they didn't with p-values!)?

Conclusion Results from the qualitative analysis revealed that, despite the frequent reporting of CIs, incidences of CI interpretation were rare. Of the articles reporting CIs, the vast majority still made their interpretations in NHST terms: They continued to make references to the null hypothesis and to discuss results in terms of *significant* and/or *non-significant*. In many ways, the discussion sections of these papers were identical to those in p-value papers. In other words, CIs had been reported (added to tables, text, and occasionally figures) to fulfill editorial hurdles, but they had made little impact on how researchers thought about and interpreted their results. The discrepancy between the proportion of reporting (the quantitative component of the study) and incidences of interpretation (the qualitative component of the study) revealed that the seemingly successful statistical reform in medicine was in fact relatively superficial.

In this study the use of mixed methods revealed a more complete picture: Medical researchers conformed to the new reporting policy and included CIs in their papers, but there had been no substantial cognitive change from dichotomous NHST thinking to CI estimation-based thinking. Fidler et al. (2004) concluded that "editors can lead researchers to confidence intervals, but can't make them think" (p. 119).

3. ACCESS TO PROCESSES AND REASONING: FAULKNER (2005)

Qualitative methods help describe complex mental processes and reasoning that are difficult to examine with quantitative methods alone. Faulkner (2005) provides an example. Faulkner aimed to improve probationary psychologists' interpretation of the outcomes of Randomized Control Trials (RCT). The study was again motivated by the argument that CIs are easier to understand than NHST, and can elicit more comprehensive and adequate interpretations (e.g., Schmidt, 1996; Schmidt & Hunter, 1997). Thirty-five probationary psychologists took part in a teaching intervention, which

consisted of one-to-one tutorials on how to interpret various RCT outcomes. In some RCT scenarios results were presented as NHST p-values and in others exactly the same results were presented as CIs. Immediately after the intervention, the participants completed two tasks. First, the participants rated their preference for each of the two presentation styles on Likert scales (quantitative). Second, they wrote short interpretations of results of some new RCT scenarios in their own words (qualitative).

Quantitative Students rated their preference for NHST or CI presentation on a 7-point Likert scale (e.g., 1=strongly prefer CI format, 4=indifferent, 7=strongly prefer NHST format). Overall, 75% of participants expressed a preference (i.e., *strongly*, *somewhat*, or *slightly* preferred) towards the CI format. Only a minority of participants (25%) had any level of preference for the NHST format.

Qualitative Students wrote short interpretations of RCT results presented as either CIs or NHST p-values in their own words. We coded and analysed their texts. In our analysis of qualitative data, we considered the comprehensiveness, structure, and quality of their descriptions.

For comprehensiveness, we looked at the number of descriptions containing the following five components: (1) the direction of effect, (2) effect size, (3) clinical significance, (4) difference between groups/statistical significance, and (5) power/precision (interval width). To analyse structure we looked at how similar each of the students' responses were. Was there a routine answer, or a lot of variation in their responses? Finally, for quality we examined whether qualifying and linking statements were used to make conceptual connections between the five components in the comprehensiveness list above.

For both NHST and CI presentations of results, students' descriptions were surprisingly comprehensive, with above 90% of students mentioning components (1) to (4). The only substantial difference between the presentation formats was in how often students mentioned (5) power/precision. When results were presented as NHST, only 70% of students made mention of power/precision; when results were presented as CIs, 97% of students did.

The analysis of structure revealed that participants generally resorted to a rigid interpretational routine when presented with NHST. CI descriptions in comparison were more varied in both content and order. Table 1 provides some typical examples of interpretations of the two formats.

As mentioned, when assessing quality we looked for qualifying and linking statements that reflected conceptual connections between the components listed above. In other words, we searched students' answers for any extra elements within the NHST and CI descriptions that were not part of the tutorial instructions. Qualifying statements included statements such as "a large effect size is good" or "clinical significance of 50% is encouraging." Examples of linking statements included "effect size is large leading to a clinically significant effect" and "non-statistically significant results were due to low power." Examples of overall conclusions included "therapy has a good effect overall" and "I would use Therapy A because it appeared to have a greater effect." On average, these extra elements were found in 90% of descriptions of CI results, compared to only 15% of descriptions of NHST results. In sum, the qualitative analysis in Faulkner's (2005) study supported the argument that CIs can elicit better, more insightful interpretations.

Table 1. Typical interpretations (verbatim) of the CI and NHST formats, for the non-statistically significant RCT scenarios (n = 35)

CI format
We can be 95% confident that Treatment X produced a large and clinically significant improvement in phobic scores. Treatment Y also showed a medium to large improvement which may or may not have been clinically significant. However (based on > ¼ overlap of CI) we cannot be sure if the two treatments had different effects. The finding for Treatment X was more precise than for Treatment Y, however, each was reasonably precise.
NHST format
There was NOT a significant interaction between treatments. There was a significant main effect for time, but not for group, meaning that scores at post-treatment were significantly lower than at pre-treatment, but the decrease was not dependent on group. There was a large effect size for both treatments. Effect was in a negative direction. 74% of the Treatment A group and 60% of the Treatment B group made clinically significant improvements, however there was not sufficient power to detect any change, HENCE, these results are meaningless (i.e., a significant interaction between treatments may exist).

Conclusion Our quantitative results revealed that students generally preferred CIs over NHST for the presentation of RCT outcomes. The analysis of qualitative responses further revealed that the CI presentations elicited interpretations that were more conceptually coherent and often included substantive interpretation of effect size; this occurred substantially less frequently with NHST presentations. As noted earlier, statistical judgments often involve complex reasoning that cannot be readily operationalised into quantitative measures. The written responses have uncovered the participants' thought processes that we were then able to break down and evaluate their structure and quality.

4. NOVEL INSIGHTS AND NEW DIRECTIONS: COULSON ET AL. (2010)

Suppose two methodologically identical studies were conducted. The difference between their observed effect sizes was small, but one yielded a p-value of .02 and the other of .22. Would researchers regard these results as consistent, or as conflicting? Coulson et al. (2010) asked this question of 330 researchers from Psychology, Behavioural Neuroscience, and Medicine. Each participant was presented with results of the two studies described above, presented either as p-values with a text description, p-values with a figure, CIs with a text description, or CIs with a figure. In each presentation the numerical results were identical and in no case should the two studies have been considered to be conflicting: The results of both vignettes overlapped considerably and a combined result of the two studies was p=.008, 95% CI: (0.81, 5.29).

Quantitative Participants were asked to rate the statement "The results of the two students are broadly consistent" using a seven-point Likert scale (1=strongly disagree, to 7=strongly agree). Coulson et al. hypothesised that the two CI formats would lead to a larger proportion of correct inferences. However, the results showed enormous variation and, to our surprise, the difference between the presentations was far from compelling: the CI advantage was only 0.67 of a point on the 7-point scale, 95% CI: (0.11, 1.23). Disappointingly, CIs did not appear to reduce dichotomous decision making. Were our predictions wrong? Was the sample skewed somehow? It was the qualitative component

of the study—where participants were asked to interpret the results in their own words—that provided the answers.

Qualitative For the qualitative component of this study participants were asked to answer in their own words, "What do you feel is the main conclusion from these studies?" A review of the text responses found that participants' interpretation of the CIs could be grouped into two categories: those who interpreted CIs as a surrogate for NHST, that is, that merely used the CI to do a hypothesis test; and those who made estimation-based responses. We coded an interpretation as an NHST surrogate if it mentioned the null hypothesis, the null value being inside or outside the interval, p-values, or statistical significance. Evidence-based interpretations of CIs, by contrast, focused on CI overlap, precision, and effect size. Of those using NHST-surrogate interpretations of CIs, 60% mistakenly identified the studies being in conflict. Of those who made estimation-based interpretations, only 5% made the error. When the dichotomous decision making procedure of NHST was eschewed, the obvious similarity of the studies was immediately apparent.

The coding rules for these classifications were developed post hoc, so some caution should be taken with these results. To rectify this, we replicated the study with the coding scheme we had developed. The replication study had 50 psychology researchers rate their agreement to the same statement used previously: "The results of the two students are broadly consistent" and again write their own interpretations of the results. Again, the text responses were coded and again our results showed that those who interpret CIs as NHST surrogates do far worse at judging the consistency of results across studies than those who make estimation-based interpretations. Of those who interpreted CIs as NHST surrogates, 92% erroneously concluded that the studies were inconsistent whereas only 21% of those who made estimation-based interpretations made this mistake. Coulson et al. (2010) concluded that CIs do lead to a better interpretation of results, but only if they are interpreted as effect size estimates. If researchers use NHST to interpret CIs, the dichotomous thinking associated with NHST will lead them astray.

Conclusion Without the qualitative component of this study we would not have explored and uncovered an integral part of researchers' statistical thinking. We found that it was not reporting format (NHST or CI) itself that improved researchers' ability to think meta-analytically, that is, to integrate results across studies. Instead, it was how these formats were interpreted. In this study the quantitative responses informed us that we were overly optimistic about how intuitively informative CIs are. On average, merely presenting results as CIs rather than NHST had little benefit. However, our qualitative analysis revealed an important distinction hidden in that quantitative summary. When a CI is appropriately interpreted as an estimation-based statistic, rather than a dichotomous NHST-based statistic, there is a marked improvement in meta-analytic thinking. This inspired a new direction for our research, which now explores a variety of ways to help students overcome dichotomous thinking.

5. TRIANGULATION: KALINOWSKI (2010)

For any given CI around a mean, is it more likely that the population mean (μ) will be near the midpoint of the CI, compared to the edges? How does the likelihood of capturing μ change as you travel along the CI arms and then outside the interval?

In most cases a *z*- or *t*-distribution is the correct way to think about these questions, with the centre of a CI relatively more likely to fall near μ than the ends of the CI. A

subjective likelihood distribution (SLD) captures a person's intuition about the relative likelihood of each point along a CI falling on the population mean (or any other population parameter being investigated). SLDs come in as many shapes as there are conceptions and misconceptions about CIs (see Figure 1 for some examples). A particularly common mistaken SLD has a uniform distribution, implying that each point in the CI is equally likely to land on μ and each point out of the interval is equally unlikely to land on μ, with a sharp cliff separating the two. In Kalinowski (2010), we investigated the prevalence of such mistaken SLDs in a student sample. Based on anecdotal experience in the classroom, we hypothesised that the SLDs in panels A and B (Figure 1) would be particularly common.

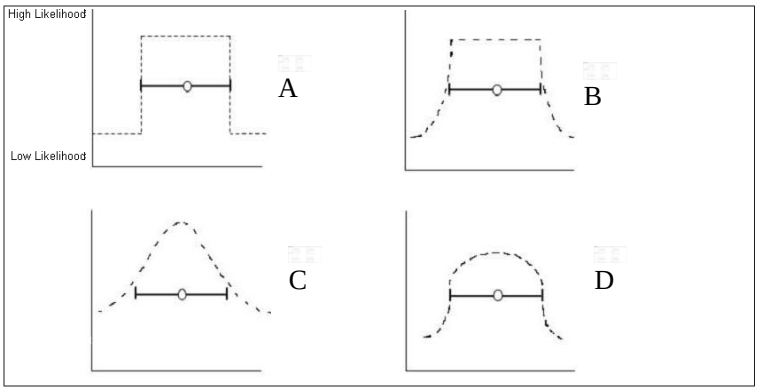

Figure 1. Some mistaken Subjective Likelihood Distributions for CIs
Panel A shows a uniform SLD, where all points in the interval are regarded as equally likely to fall on μ and all points outside the interval are equally unlikely to fall on μ, with a sharp cliff separating the two. Panels B to D show other SLDs.

Quantitative We measured students' SLDs by presenting students with a CI figure (Figure 2. Pictured is the figure shown for a 95% CI. Participants were also shown a 50% CI.) and asking them to rate the likelihood of each labeled point relative to the previous labeled point, on a 13 point rating scale (1=much less likely, 7=equally likely, 13=much more likely). We chose a 13 point rating scale, rather a 7- or 9-point scale, to achieve finer mapping when plotting SLDs from these ratings. Analysis of participant responses showed that 27%, 95% CI: (19, 37), of students gave answers consistent with SLD A in Figure 1 and 13% (8, 22) gave responses consistent with SLD B. Together these two SLDs accounted for 40% of our student sample.

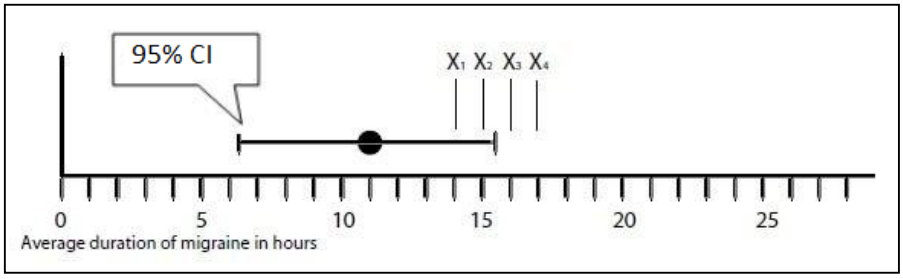

Figure 2. Participants were asked to rate the likelihood of each point (i.e., X_2) relative to the previous point (i.e., X_1)

Qualitative In the qualitative component of this study students were asked, "How did you go about answering these questions? Did you have any particular model in mind? Did you use any rules of thumb?" Students wrote open-ended responses in their own words. Some of the answers were uninformative or vague, suggesting that some responses were somewhat random. Others, however, gave clear and concise explanations for their reasoning. For example, the quotations below are unambiguous endorsements of a uniform SLD:

Graduate Student #8:	I answered these questions believing that for a 95% CI there is a 95% chance that the true population mean is within those parameters. I did not think the likelihood would be affected by a value's distance to the mean as long as the values were within the 95% CI.
Graduate Student #32:	The true population mean is equally likely within the confidence interval regardless of distance from the sample mean.
Graduate Student #50:	To the best of my understanding, a confidence interval does not relate any information as to the probability that a particular value within the confidence interval being any more or less likely to be the population mean.

Conclusion In the quantitative component we had made our best guess at how to operationalise the abstract concept of SLDs, using a rating scale for various points in and out of the CI. We did not know whether the scale was an adequate way of capturing the SLD concept; that is, we did not know that what we were eliciting on the scale was in fact a representation of an SLD. Our qualitative responses provided assurance that our scale was measuring what it was supposed to. In this sense, the qualitative response helped validate our quantitative measure. Furthermore, in this study two measures afforded triangulation: Both quantitative and qualitative responses confirmed that a mistaken uniform SLD is highly prevalent in the student population.

6. ENSURING ROBUSTNESS IN QUALITATIVE RESEARCH

Qualitative methods are often mis-associated with terms such as *subjective* or *biased*. We believe that subjectivity itself is not necessarily the problem in either qualitative or quantitative research. Rather, it is disclosure of subjectivity and potential biases, and the transparency of uncertainty, that is crucial to remaining scientific. Research judgement is an integral and important part of both quantitative and qualitative methods. Similarly, bias can be present in all research, and both quantitative and qualitative research can be done well or poorly. Like qualitative researchers, quantitative researchers have their own preconceptions and biases (e.g., A and B do *significantly* differ). These expectations are often hidden in seemingly neutral null hypotheses of *no difference*, even though the motivation for doing the research must be a reasonably strong belief in the opposite! Researchers are expected to design an experiment in such a way that would "rarely fail to give us a statistical significant result" (Fisher, 1960, p. 14). Yet, statistical power is often low and unreported. The illusion of certainty (*significant* or *non-significant*) is maintained by the dichotomy of NHST, but in fact there is little objectivity in the procedure, which has repeatedly shown to be subject to a string of fallacies and misconceptions (Kline, 2004). These criticisms of the typical approach to quantitative methods do not in themselves constitute an argument for qualitative methods. Rather, we present them here

only to highlight the fact that there is no *a priori* sense in which one or the other method is more objective.

Elliott et al. (1999) proposed a set of seven guidelines for qualitative research: 1. Own your perspective, 2. Situate the sample, 3. Ground in examples, 4. Provide credibility checks, 5. Ensure coherence, 6. Accomplishing general vs. specific research tasks, and 7. Resonate with the readers. Below we explain in our terms what each of these guidelines (as well as *triangulation*) means. We briefly illustrate how we have endeavoured to meet them in the example studies discussed above. Finally, we argue that these guidelines should, in most cases, also apply to quantitative research.

6.1. OWNING ONE'S OWN PERSPECTIVE

Owning your own perspective involves acknowledging theoretical predispositions and expectations. The purpose of owning one's perspective is to allow the reader to critically review the researchers' position, method, data, and interpretation. It also invites the reader to consider other interpretations of the data when different relevant theoretical positions apply. Writing in the first person, rather than the anonymous third person, assists in achieving the aim of "owning your own perspective." We made a disclosure of this kind in the opening paragraph of this paper, by stating our position on NHST and statistical reform.

6.2. SITUATING THE SAMPLE

Situating the sample involves describing the participants in detail. Sufficient detail will depend on the context of the study but may include basic demographics (e.g., age, gender) as well as the participants' theoretical position, their motivation for participation, and their attitude to the study (e.g., engagement in the interview process). In our studies, important participant details include the level of training in statistics and their general attitudes toward statistics. In Coulson et al. (2010), for example, we expected that our final sample was self-selected (recall that we emailed authors of published articles in leading journals). Our original sampling procedure was reasonable, but of course, not all the emails we sent elicited replies. It may be reasonable to assume that those who replied were more confident, competent, or interested in statistics than those who declined our invitation to participate. If our sampling method was biased in this way, then the results are interesting in that they show a high rate of misunderstanding even in this expert group. However, a competing hypothesis is that only researchers who were less busy replied. If this is the case, then the interpretation of our results might be quite different.

6.3. GROUNDING IN EXAMPLES

Grounding in examples means to give clear and detailed examples of the questions being asked and the responses given. We hope we followed this guideline throughout by quoting research questions used and explaining response scales and tasks. In the Faulkner (2005) and Kalinowski (2010) studies we also provide verbatim examples of responses. Again these examples invite the reader to appraise our interpretations and think about other ways the data could have been interpreted. We believe using examples is also important in quantitative research; for example, we always provide the exact wording of questions asked for quantitative components of our research as well.

6.4. CREDIBILITY CHECKS

For the types of studies given in our examples, credibility checks most often involve cross coding. The ideal case is to include someone unfamiliar with the research program to perform this procedure (blinded cross-coding). However, basic understanding about the background materials is almost always required in order to make sense of important terms and concepts. A detailed coding manual is therefore essential. A coding manual outlines the categories to be coded and offers definitions and usually examples. Its goal is to eliminate linguistic ambiguity as much as possible. A good coding manual should allow someone with only minimal training to repeat the coding activity. Depending on the size of the coding activity, double coding all items may be possible. For larger projects this is usually impractical, and so a sample of between 10% and 20% of items (e.g., students' open-ended survey responses, interview transcripts) will be cross-coded. The agreement between coders is then calculated. This can be done in several ways. A percentage agreement is the most basic and often adequate. Cohen's (1960) Kappa may be useful in cases where it is important to account for agreement occurring by chance. We usually consider agreement of around 80% to be adequate and 90% to be good, although we acknowledge that this decision is arbitrary. The operationalisation of credibility checks will be different in quantitative research, but the concept remains of crucial importance.

6.5. COHERENCE

Coherence refers to logical organisation and presentation of argument and evidence. It also involves explaining the logical interconnectedness of separate pieces of evidence. In the Conclusion section of each of our four studies we have attempted to present a summary that highlights the connections between our quantitative evidence and our qualitative evidence, and, where necessary, explains discrepancies. Again, we believe coherence of argument and evidence should be equally important in both qualitative and quantitative research.

6.6. ACCOMPLISHING GENERAL VS. SPECIFIC RESEARCH GOALS

Some studies may have quite specific goals of describing a particular sample or group. Others wish to use a sample or group to make general claims across time, people, and/or place. Identifying the goal of any given research project is crucially important in both qualitative and quantitative research. In quantitative research this usually involves stating whether the research is descriptive or inferential. Many quantitative reports may claim to be descriptive, but inferential goals sneak into the discussion. Similarly, many qualitative research reports begin by stating a specific research goal, but end up wanting to fulfill a general goal.

An aside on generalisability and replicability is relevant here. Confirmation of a theory comes from the repetition of experiments, which increases researchers' confidence about the adequacy of the theory, and the predictions derived for the studied and associated phenomena (Tukey, 1969). For example, in the Coulson et al. (2010) study we repeated the experiment after developing a coding guide and refining our hypothesis. Research repetition is a crucial activity in science for both confirmation of theories and validation of previous findings. Although not equivalent, generalisability and replicability share a close relationship. For example, if the sampling method is adequate, then findings of a study (both qualitative and quantitative) can readily be generalized to the population from which the sample was drawn. In reality, replications inevitably involve some

methodological changes (i.e., new experimenters, samples, and questions). These changes can test robustness, or generalisability, of findings over a broader range of populations and conditions. The narrow view of a replicated experiment is that a result is obtained again, in a "pure" replication experiment; that is, one that's identical except with a new sample. That sort of replication reduces the role of sampling fluctuations, so we can be more confident of the result. It increases precision. In qualitative research, replicability can be facilitated by a detailed documentation of coding forms and researcher instructions in the method (or appendix) section of a paper.

6.7. RESONATING WITH THE READER

Following the above guidelines should help ensure that the final manuscript resonates with the reader: Disclosing biases, avoiding ambiguity, being transparent, writing in the first person, grounding in examples, providing credibility checks, and ensuring coherence are all important steps in ensuring resonance. In addition, writing clearly and engagingly is always a fine goal. It goes without saying that this advice applies to quantitative research equally.

6.8. TRIANGULATION

Triangulation is reaching the same answer through different means. It helps to demonstrate that the phenomenon that has been observed is not merely a product of the instrument or method used to make the observation. We described an example of triangulation in the Kalinowski (2010) study, where we found both quantitative and qualitative evidence for the existence of misconceived uniform SLDs about CIs.

7. CONCLUSION

In this paper we have proposed and given examples of reasons why we believe qualitative research is essential to the study of statistical cognition. To summarise again, these were to provide more complete description of phenomena, for access to processes and reasoning, to uncover novel insights and new directions for research, and triangulation.

In addition, qualitative data can explain an incorrect interpretation of quantitative data (Coulson et al., 2010) or further support interpretations (Kalinowski, 2010). Methodological rigour of qualitative research is possible and there are guidelines available—such as those proposed by Elliott et al. (1999)—to help qualitative research maintain robustness. We have further argued that these same guidelines can be used to strengthen quantitative research. When researchers maintain transparency and avoid ambiguity throughout the experimental process, analysis of results, and write-up, readers are able to evaluate the evidence themselves and judge the validity of the interpretation. Transparency also allows readers to create a faithful replication of the research. Detailed coding instructions, coder training, and theoretically justifiable coding categories are particularly crucial to this process.

Finally, both quantitative and qualitative methods are to some extent biased, and both quantitative and qualitative research can be done well or poorly. Unfortunately, it seems that only qualitative researchers have been required to state plainly what their biases actually are. Quantitative research has for too long hidden behind a wall of seemingly objective statistical hypothesis without ensuring the same levels of scrutiny as their

qualitative peers. Perhaps this is something that quantitative research could take from its lesser well known, and commonly mistrusted, counterpart?

We believe using mixed methods limits the inherent shortcomings of both qualitative and quantitative research methods. It is especially useful when there is interest in both process and outcome, which is the case in our statistical cognition research program. We want to understand students' and researchers' reasoning, but we also want to test whether and how much intervention, new displays, editorial policy, and teaching methods can improve reasoning. For example, does providing researchers with CIs rather than p-values facilitate better interpretation? Does enforcing strict editorial policy improve estimation-based thinking? Using qualitative research methods helps us identify and access the underlying cognitive processes, and quantitative research methods allow us to corroborate, generalise, and test the outcomes of those processes. A mixed methods approach makes our research more complete and worthwhile. Qualitative research without a quantitative component can be insightful but vague; quantitative research without a qualitative component can be precise but may be wrong-headed, misdirected, and contrived.

REFERENCES

Beyth-Maron, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7*(2), 20–39.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement, 1*(26), 1–9.
[Online:www.frontiersin.org/quantitative_psychology_and_measurement/10.3389/fpsyg.2010.00026/abstract]

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

Elliott, R., Fischer, C., & Rennie, D. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology, 38*(3), 215–229.

Faulkner, C. (2005). *Randomized controlled trials in clinical psychology: Towards better understanding of research results using confidence intervals and other statistics* (Unpublished doctoral dissertation). La Trobe University, Melbourne, Australia.

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., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*(2), 119–126.

Fisher, R. A. (1960). *The design of experiments* (7th ed.). New York: Hafner.

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. In C. Reading (Ed.), *ICOTS-8 Proceedings: Towards an evidence based society.* Voorburg, The Netherlands: International Association for Statistical Education, International Statistics Institute.
[Online: http://icots8.org/cd/pdfs/contributed/ICOTS8_C104_KALINOWSKI.pdf]

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L, L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in *The Journal of Experimental Education. The Journal of Experimental Education*, *66*(1), 75–83.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83–91.

PAV KALINOWSKI
La Trobe University Victoria, 3086
Australia