April McGrath, MOUNT ROYAL UNIVERSITY, amcgrath@mtroyal.ca

Searching for Significance in the Scholarship of Teaching and Learning and Finding None: Understanding Non-Significant Results

ABSTRACT

Quantitative results from empirical studies are common in the field of Scholarship of Teaching and Learning (SoTL), but it is important to remain aware of what the results from our studies can, and cannot, tell us. Oftentimes studies conducted to examine teaching and learning are constrained by class size. Small sample sizes negatively influence statistical power and make non-significant results a more likely occurrence. When one finds non-significant results it is important to consider what conclusions can be drawn from the study. This article provides information on null hypothesis significance testing that is relevant to our understanding of non-significant results, and it highlights the importance of recognizing underpowered studies in the teaching and learning literature. Factors that can contribute to non-significant findings in a study are also highlighted. Being aware of these factors, statistical power, and the logic of significance testing will put scholars in a better position to evaluate non-significant results from their own research and that of others.

KEYWORDS

null hypothesis significance testing, statistical power, non-significant effect

INTRODUCTION

Interest in Scholarship of Teaching and Learning (SoTL) continues to grow, and instructors wishing to disseminate findings from investigations about teaching and learning in their classrooms have a number of outlets. With this growth have come calls for clear standards to evaluate the research designs and analyses of SoTL (Wilson-Doenges & Gurung, 2013). In particular, for scholars interested in conducting or consuming empirical studies of teaching and learning that rely on quantitative analyses, a familiarity with statistics is key. And all scholars, even those with training in statistics, require reminders, or clarification, from time to time about particular methodological and analytical stumbling blocks. This article focuses on one specific stumbling block—results that do not achieve statistical significance—and the confusion surrounding the interpretation of such results. For SoTL to be of maximal usefulness, a proper understanding of statistically non-significant effects is necessary among both researchers and readers.

Before detailing the confusion around statistically non-significant effects, it is important to keep in mind what a statistically significant difference actually means. Statistically significant differences are thought to demonstrate real effects, as opposed to differences resulting purely from sampling error. When an author reports a test statistic followed by $p < .05$, the proper interpretation of this result is that the obtained difference, or a larger difference, would only be observed less than 5% of the time if the null hypothesis were true. Another way to understand this is to think *if there is really no difference between populations* (i.e., if the null hypothesis is true) then this obtained result was highly unlikely. In fact, this

result is so unlikely (less than a probability of .05) I reject the initial assumption of no difference between populations (i.e., reject the null hypothesis) and conclude instead that a difference does exist. In a nutshell, that is the logic of null hypothesis significance testing (NHST), which is a dominant statistical approach in many fields. Scholars are often interested in demonstrating the influence of a technique, or the existence of a relationship, and as such would design a study with an aim to reject the null hypothesis. When all goes according to plan, scholars observe data that support the existence of an effect and publish the findings. However, studies can also produce data that do not support rejection of the null hypothesis. When this is the case, what can one conclude from a study? Answering this question in the context of SoTL is the purpose of the present paper.

The difficulty of interpreting non-significant effects is, in part, highlighted by a recent study exploring bottleneck concepts within psychology (Gurung & Landrum, 2013). In an effort to identify bottleneck concepts for students, a sample of instructors rated the difficulty of 91 terms from psychology. The third most difficult item on the list was *statistical significance*. Relatedly, a sample of students was asked to assess the difficulty of, and their perceived confidence in, understanding 60 terms from research methods. Bottleneck scores were identified by the pairing of a high difficulty score with a high confidence score. Among this list, *null hypothesis* received the fifth highest bottleneck score. Thus, results suggest that faculty suspect students will have difficulty understanding the *statistical significance* from a hypothesis test, and students, while aware of the difficulty of the related term *null hypothesis*, nonetheless tend to overestimate their ability to understand it. Together these findings suggest concepts about hypothesis testing pose challenges for students, but it is not just students who struggle with statistical reasoning. Studies with samples of psychological researchers have noted important limitations in statistical reasoning (Tversky & Kahneman, 1971; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Scholars have identified a number of ways that the logic of NHST and the meaning of p values are misinterpreted. Two excellent reviews of such misinterpretations are provided by Kline (2004) and Goodman (2008). Below we will concern ourselves with one particular sticking point—the interpretation of non-significant findings; that is, failing to reject the null hypothesis.

Having a clear understanding of non-significant results in SoTL is important so that we can properly evaluate the learning outcomes from research studies. As an example, consider the work of Saville, Pope, Lovaas, and Williams (2012) concerning interteaching and the use of post-discussion quizzes. Interteaching involves the use of preparation guides. Students read these guides and answer items from the guides before class. In-class time is devoted to a clarifying lecture at the beginning about previous material from a preparation guide that was particularly difficult for students to understand. Students then work in pairs to discuss items from the preparation guide, and difficult items will serve as the basis for the next class's clarifying lecture (Saville, 2013). Positive learning outcomes from this approach have been documented (Saville, Zinn, & Elliott, 2005; Saville, Zinn, Neef, Van Norman, & Ferreri, 2006). In this particular study (Saville et al., 2012), the researchers paired the use of interteaching with an intervention designed to create the testing effect. The testing effect refers to the well-documented positive learning gains that are observed when students retrieve information from memory (Roediger & Butler, 2011). The researchers sought to further improve the effectiveness of interteaching by incorporating this additional component. Despite students completing post-discussion quizzes prior to some exams, these students did not outperform peers serving as a control group at a statistically significant level. The authors have suggested these findings (i.e., non-significant results) provide evidence that post-discussion quizzes may not enhance learning in an interteaching classroom.

This is a classic example of the difficulty we are presented with when interpreting results from hypothesis testing. Before concluding that a non-significant result provides evidence of no treatment effect, we must consider logically what a non-significant finding can tell us and we must also consider the role of statistical power in our research designs, especially in SoTL.

Interpreting a non-significant effect makes for difficult decision-making. In describing NHST, some authors have taken care to warn researchers about the conclusions one can draw from such effects. For example, in their introductory text Aron, Coups, and Aron (2013, p. 115) note that a non-significant effect essentially leads to an inconclusive finding: "the results may not be extreme enough to reject the null hypothesis, but the null hypothesis might still be false". In a recent evaluation of the statistics commonly reported across several major journals, important limitations of NHST were outlined, one of which directly relates to the interpretation of non-significant effects (Tressoldi, Giofré, Sella, & Cumming, 2013). Using stronger language, these authors noted it is a *logical fallacy* to conclude that a null hypothesis is true in an instance where one fails to reject the null. Again, it is best to view this as an inconclusive finding:

> *Remember that the null hypothesis is that there is no effect in the population. All that a non-significant result tells us is that the effect is not big enough to be anything other than a chance finding—it doesn't tell us that the effect is zero. As Cohen (1990) points out, a non-significant result should never be interpreted (despite the fact it often is) as 'no difference between means' or 'no relationship between variables.'* (Field, 2009, p. 53)

Kline (2004, p. 67) also emphasizes this point by reminding readers as a "basic tenet of science that absence of evidence is not evidence of absence". That is, just because a difference is not statistically significant does not mean that a difference does not exist. It is tempting to think that a non-significant finding means we should accept the null hypothesis because NHST may encourage a researcher to see decisions about results as black and white rather than gray. Cumming (2014) notes that the dichotomous thinking encouraged by NHST is a major challenge to overcome to have a better appreciation for what statistical findings are telling and not telling us.

Beyond the fact that logically a non-significant result cannot translate into acceptance of the null hypothesis, another important factor to consider in our interpretation of non-significant effects is statistical power. Statistical power is the likelihood of detecting an effect that exists. If a study lacks sufficient statistical power the probability of a Type II error is more likely. Type II errors are instances where the researcher did not reject the null hypothesis even though the null hypothesis should be rejected. It is akin to not finding a statistically significant effect, even though the effect actually exists. The chance of a Type II error is particularly important to keep in mind for studies about teaching and learning that are restricted to small classes and thus small sample sizes. This is because statistical power increases as sample size increases. Recently, Tomcho and Foels (2009) reviewed the power of statistical tests from 197 studies published in *Teaching of Psychology*. Their findings provide an important reminder about interpreting non-significant effects. They found that many published studies were underpowered. That is, sample sizes were typically too small and researchers often failed to account for the detection of medium or small effects rather than large effects. These findings are of particular relevance to the interpretation of non-significant effects in the teaching and learning literature where sample size is often constrained by course enrollment. Smaller sample sizes mean less statistical power. Before jumping to conclusions about a seemingly failed teaching intervention, we need to assess whether

the study had a large enough sample to detect an effect. Studies with non-significant results and small sample sizes hardly provide evidence of failed teaching interventions.

In the interteaching study noted above, Saville et al. (2012) used an independent groups design with a total of 58 participants across two conditions. According to Cohen (1992) such a sample size would only provide the recommended level of power (.80) in situations where large treatment effects are present. As such, to conclude that the post-discussion quiz intervention may not enhance learning is premature. Perhaps a better explanation of the findings rests on a lack of statistical power, especially given that five of the six statistical tests demonstrated non-significant differences in the predicted direction. As such, scholars need to be aware of statistical power in order to properly interpret results that may at first suggest a failed teaching intervention.

Just as non-significant results are at times confused as evidence of zero treatment effect, such results are also used at times to suggest group equivalency prior to the start of an intervention (see misconception #2 as identified by Goodman, 2008). In studies about teaching and learning, it is often not possible to assess learning outcomes through the use of randomly assigned groups. In situations where random assignment is not possible, scholars still want to demonstrate that the comparison groups are similar at the beginning of the study. Unfortunately, comparisons between groups on several different variables that result in non-significant differences do not mean that the groups are equivalent to begin with. This is not to suggest that we should not strive for appropriate comparison groups and groups that are similar to one another. But it does mean that we should avoid accepting the false comfort offered by a claim that groups are equivalent simply because no statistically significant differences were detected between them. As noted by Goodman (2008, p. 136), "a nonsignificant difference merely means that a null effect is statistically consistent with the observed results, together with the range of effects included in the confidence interval (CI). It does not make the null effect the most likely". Furthermore, studies with small sample sizes will lower statistical power, thus making the occurrence of non-significant results from pre-intervention group comparisons more likely.

Grauerholz and Main (2013) also warn against assuming that groups are equivalent when studying teaching and learning. They note that pre-existing differences between classes as well as differences that develop over the course of the term would lead one to doubt that groups based on class enrollment are equivalent. For example, despite having two sections of one course that were comparable in gender ratio, size, delivery time, and background, Grauerholz notes that developments in the classes beginning on day one made them unique. As such, comparisons between the groups resulting in non-significant differences should not lead one to think the groups are equivalent. Although striving for high standards in our research designs is important (e.g., using random assignment when possible and appropriate, using appropriate comparison groups), we should not let a *desire* for rigor lead us astray by incorrectly interpreting statistical findings to suggest something they do not. Again, absence of evidence is not evidence of absence. Non-significant differences between groups pre-study do not mean the groups are equivalent.

Given the preceding discussion it may seem that a non-significant effect leaves one at a dead end or perhaps in limbo. This is not the case. Going beyond the calculation of a *p* value, and also including a power analysis, confidence intervals, and measures of effect size will give scholars greater understanding about the data that comes from studies about teaching and learning (see Cumming 2014). Ultimately, we need to do our best to ensure conclusion validity, which is ascertaining reasonable conclusions about relationships based on our data (Trochim, 2006). Non-significant effects require careful consideration and rushing to conclusions about no effects or no relationships will only impede progress in SoTL.

Consequently, a substantial interrogation of non-significant effects is important. Morling (2015) provides an accessible and excellent summary of points to consider when confronted with a non-significant effect. She lists possible explanations for such findings and places them in two categories: insufficient between-groups difference and too much within-groups difference. Insufficient between-groups difference refers to a failure to observe notable differences between groups, perhaps between a control and experimental group. On the other hand, a failure to observe a statistically significant effect could be due to sizeable within-groups difference, also known as noise. To briefly summarize, she reminds readers to evaluate non-significant effects by considering 1) the strength of the manipulation of the independent variable. For example, if I believe feedback influences the quality of students' writing I could devise a study where one group of students receives feedback and the other does not (the ethical implications of this study are not considered here). However, if my feedback consists of merely one short sentence it is possible that I will not observe differences in the quality of work between these groups because my intervention was not strong enough. A relationship between feedback and writing quality may exist, but my intervention was not strong enough to demonstrate the relationship. In evaluating findings we must also consider 2) the sensitivity of the dependent variable. Again, in the context of a feedback and writing study, if the student papers were evaluated on a pass fail basis, rather than a numerical scale, I may observe a non-significant difference between the groups simply because my outcome measure was too insensitive to track the differences in the learning outcome. Also, we need to remember 3) the potential of measurement error, individual differences, and situation noise to increase unsystematic variability thereby obscuring a treatment effect (p. 326). In the context of a feedback and writing study, inaccurate and unreliable measures of writing ability, individual differences between students in writing ability, and outside influences such as the availability of writing tutors or informal peer review activities could all contribute to within group variability that would obscure between group variability and ultimately the detection of a statistically significant effect. By assessing studies with non-significant findings against the above criteria we can determine whether a good effort was put forth to find an effect, which is a key consideration outlined by Frick (1995) in his discussion about accepting the null hypothesis.

Only after considering these factors can one *perhaps* begin to consider that no relationship exists between the independent and dependent variables. In terms of understanding non-significant effects within SoTL, I think we should keep in mind these factors while also remembering the important role of sample size and, consequently, statistical power.

To summarize, before concluding a teaching activity is ineffective, researchers should ask themselves the questions posed by Morling (2015). Prior to conducting a study, researchers should plan for the appropriate sample size to achieve 80% power (statistical power calculators can be found online). Specific recommendations on ways to increase power are provided by Tomcho and Foels (2009) and readers may also wish to consult Cohen's (1988) work on the topic. Lastly, Tomcho and Foels (2009) recommend that researchers report statistical power to aid the reader's interpretation of results, and I believe this is an important consideration to further the work of our field. Providing more statistical information in our articles (e.g., power, effect size, CIs) and being cognizant of what findings from NHST can and cannot tell us will lead to more robust conversations and investigations about teaching and learning.

April McGrath is an Associate Professor in the Department of Psychology at Mount Royal University. She was a 2012 Nexen Scholar at the Institute for Scholarship of Teaching and Learning.

McGrath, A. (2016). Searching for significance in the scholarship of teaching and learning and finding none: Understanding non-significant results. *Teaching & Learning Inquiry, 4*(2). http://dx.doi.org/10.20343/teachlearninqu.4.2.12

REFERENCES

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th ed.). New York, NY: Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29. doi: 10.1177/0956797613504966

Field, A. P. (2009). *Discovering statistics using SPSS: (And sex and drugs and rock 'n' roll)*. Thousand Oaks, California: SAGE Publications.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 25*, 132-138.

Goodman, S. (2008). A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology, 45,* 135-140. doi: 10.1053/j.seminhematol.2008.04.003

Grauerholz, L., & Main, E. (2013). Fallacies of SOTL: Rethinking how we conduct our research. In K. McKinney (Ed.), *The scholarship of teaching and learning in and across the disciplines* (pp. 152-168). Bloomington, Indiana: Indiana University Press.

Gurung, R. A. R., & Landrum, R. E. (2013). Bottleneck concepts in psychology: Exploratory first steps. *Psychology Learning and Teaching, 12*(3), 236-245.

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

Morling, B. (2015). *Research methods in psychology: Evaluating a world of information*. New York, NY: W. W. Norton & Company, Inc.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science, 15*, 20-27. doi: 10.1016/j.tics.2010.09.003

Saville, B. K. (2013, February). Interteaching: Ten tips for effective implementation. *Observer, 26*(2).

Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012). Interteaching and the testing effect: A systematic replication. *Teaching of Psychology, 39*, 280-283. doi: 10.1177/00986283124 56628

Saville, B. K., Zinn, T. E., & Elliott, M. P. (2005). Interteaching vs. traditional methods of instruction: A preliminary analysis. *Teaching of Psychology, 32*, 161-163. doi: 10.1901 /jaba.2006.42-05

Saville, B. K., Zinn, T. E., Neef, N. A., Van Norman, R., & Ferreri, S. J. (2006). A comparison of interteaching and lecture in the college classroom. *Journal of Applied Behavior Analysis, 39*, 49-61. doi: 10.1901/jaba.2009.42-369

Tomcho, T. J., & Foels, R. (2009). The power of teaching activities: Statistical and methodological recommendations. *Teaching of Psychology, 36*, 96-101. doi: 10.1080/00986280902739743

Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. (2013). High impact = high statistical standards? Not necessarily so. *PLoS ONE, 8*, 1-7. doi: 10.1371/journal.pone.0056180

Trochim, W. M. (2006). Conclusion validity. *The research methods knowledge base* (2nd ed.).

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Wilson-Doenges, G., & Gurung, R. A. R. (2013). Benchmarks for scholarly investigations of teaching and learning. *Australian Journal of Psychology, 65*, 63-70. doi: 10.1111/ajpy.12011

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.