

Journal for Articles in Support of the Null Hypothesis
Vol. 1, No. 3
Copyright 2002 by Reysen Group. 1539-8714
www.jasn.com

Interpreting Null Results: Improving Presentation and Conclusions with Confidence Intervals¹

Chris Aberson

Humboldt State University

In this paper, I present suggestions for improving the presentation of null results. Presenting results that “support” a null hypothesis requires more detailed statistical reporting than do results that reject the null hypothesis. Additionally, a change in thinking is required. Null hypothesis significance testing do not allow for conclusions about the likelihood that the null hypothesis is true, only whether it is unlikely that null is true. Use of confidence intervals around parameters such as the differences between means and effect sizes allows for conclusions about how far the population parameter could reasonably deviate from the value in the null hypothesis. In this context, reporting confidence intervals allows for stronger conclusions about the viability of the null hypothesis than does reporting of null hypothesis test statistics, probabilities, and effect sizes.

This paper discusses presentation of null results derived from traditional null hypothesis significance testing (NHST) procedures, and presents examples and suggestions for improvement in statistical reporting.² This paper does not include a complete discussion of all the issues associated with NHST; rather it is a primer with practical suggestions for data reporting. In the reference section, I include suggestions for more detailed reading on topics discussed in this paper. Left aside are issues such as whether it is reasonable to assume that the null is ever true (e.g., Cohen, 1994), Bayesian techniques (Pruzek, 1997; Krueger, 2001), three outcome tests (e.g., Harris, 1997), power for research design (e.g., Cohen 1992), graphical presentation (e.g., Loftus, 2002), meta-analytic thinking (Thompson, 2002), and countless other issues. These are relevant and important issues, however, this paper focuses exclusively on null findings resulting from traditional NHST procedures and how confidence interval presentation improves conclusions. For the sake of brevity, I focus on presentation of two-group comparisons. I present these issues in the context of providing support for a null hypothesis, however, these suggestions are relevant to presentation any statistical result.

Some Definitions

Before discussing reporting null results, a discussion of terminology is needed. Terms used throughout this paper are null hypothesis, parameter, Type I error, Type II error, power, rejecting the null hypothesis, and failing to reject the null hypothesis. The null hypothesis (H_0) is a statement of no effect. For the comparison of the means of two groups, (H_0) states that the difference between the groups in the population is zero. This relationships is commonly noted as $H_0: \mu_1 - \mu_2 = 0$, where μ_1 is the mean score on the dependent measure for group 1 and μ_2 is the mean score on the dependent measure for group 2.

¹ Stephen Reysen served as action editor for this blind reviewed article. The article's author serves as editor of JASNH but was not involved in the editorial decision involving this work.

² I prefer the terms null findings or null results to null effect as null effect suggests that there is no effect present whereas null findings or null results suggest that no effect was detected but may exist.

Technically, the null hypothesis specifies the expected value of a population parameter. Parameter refers to a population characteristic. NHST relies on sample results to draw conclusions about the value of the population parameter. Specifically, NHST focuses on the probability of obtaining an observed sample outcome given that H_0 is true. Less technically, NHST asks whether the sample result would be surprising if the two groups do not differ in the population.

Type I error is the probability of rejecting H_0 when it is true. Rejecting H_0 in this case would lead to the conclusion that the groups differ when they do not actually differ. This is sometimes referred to as alpha or the significance level. For the purpose of this paper, I will assume alpha = .05. Type II error is the probability of failing to reject H_0 when it is false. Here, the conclusion would be that the groups do not differ when they actually do differ. Power is 1 minus the Type II error.

Rejecting H_0 yields a conclusion that provides some probabilistic certainty. Assuming a Type I probability of 5% (alpha = .05), rejecting H_0 means there is only a small chance that the obtained result would have occurred if H_0 were in fact true.

What is a null result?

A null result occurs when we fail to reject H_0 . This is commonly referred to as a non-significant result or ns. There are two possible “realities” when the null is not rejected. First, is the case where the null is true. If H_0 were true, the probability of failing to reject H_0 is .95. When the null is true, it is unlikely to make a Type I error. Second, is the situation wherein H_0 is false but we fail to reject H_0 . This reflects the Type II error probability (1 - Power). If the null is actually false, this probability is ideally about .20. This means that, given an ideal situation, when H_0 is false it will not be rejected 1 out of 5 times. This result does not convey statistical certainty. It is likely that either H_0 is true or H_0 is false. Neither result is ruled unlikely. This differs sharply from situations wherein we can reject H_0 . When we reject H_0 , there is a comparatively small probability that we have made an error (i.e., Type I error rate of 5%).

Adding this problem is the relatively low statistical power in psychological research, typically between .40 and .60 (Lipsey & Wilson, 1993; Sedlmeier & Gigerenzer, 1989). A complete discussion of factors influencing power is somewhat outside of the scope of this article but the interested reader should refer to Cohen (1992 for a short overview; 1998 for a detailed discussion). If we fail to reject H_0 with 50% power, there is a high probability that the research design was not sensitive enough to detect effects. Given these issues with power, when we fail to reject the null hypothesis, we must be careful as to what we conclude as failing to reject H_0 does not tell us enough to conclude anything about the viability of H_0 .

In summary, when H_0 is true, it is likely that we will fail to reject it. When H_0 is false, we may also fail to reject H_0 due to low statistical power. In both cases, our conclusion is to fail to reject the null hypothesis (a null result). When we fail to reject the null hypothesis, it does not transmit any meaningful information about the viability of the null hypothesis, primarily because of the high probability of making a Type II error.

Can you accept the null hypothesis?

Technically no. Practically, yes. Support for the null hypothesis (e.g., accepting the null) is a tenuous proposition.³ A commonly held misconception is that failing to reject H_0 suggests that H_0 is true (Nickerson, 2000). This is not the case. Failure to reject H_0 suggests that either H_0 is true, reflecting a correct decision or H_0 is false but we do not have enough power to reject H_0 .

For example, imagine we were testing the null hypothesis that mean performances of two groups were equal (e.g., difference between groups in the population is zero or $H_0: \mu_1 - \mu_2 = 0$) and our sample result suggested that H_0 should be rejected (e.g., groups differed by 10 points, $p < .05$). This result tells us that it is unlikely that the differences in the population are zero. The result does not suggest that the actual differences in the population are 10 points, only that $\mu_1 - \mu_2$ is likely greater than 0. Rejecting H_0 is strictly a conclusion about the viability of our null hypothesis (i.e., what the population parameter is not). Rejecting the null hypothesis does not tell us what the value of the parameter is likely to be, rather, only what it is not likely to be. NHST is good at telling us what population parameter is unlikely. NHST does not tell us what the parameter is likely to be.

Now, consider an example wherein the researcher tests the same null hypothesis but the sample result suggested that H_0 should not be rejected (e.g., groups differed by 0.5 points, $p = .60$). This result tells us that it would not be surprising to find two groups that differed by this much if the parameters specified in H_0 were true. However, there are other plausible values for the parameter (e.g., $\mu_1 - \mu_2 = 1.0$). Failing to reject H_0 does not rule out any potential values for the $\mu_1 - \mu_2$, rather it suggests that a certain value (zero in this case) would not be surprising.

The two results above lead to different conclusions. Rejecting H_0 tells us something meaningful. Specifically, that the parameter specified in H_0 is not likely to be the parameter actually observed in the population. Failing to reject H_0 tells us we cannot rule out the value specified in H_0 as a likely value for the parameter. Keeping in mind that scientific reasoning centers on principles of falsification, it becomes clear that rejecting H_0 provides falsification whereas failing to reject H_0 does not. Using this reasoning, rejecting H_0 is valid scientific evidence whereas failing to reject H_0 is not.

³ Failure to reject the null with power of 95% (i.e., a 5% chance of a Type II error) provides support for the null, just as rejecting the null with a Type I probability of .05 supports rejection of the null. This is true theoretically. Practically, it is not a situation that a researcher would likely encounter.

Failure to reject H_0 should not suggest that we cannot draw meaningful conclusions or that the research is without value. Keep in mind that “proving” the null to be true is simply not viable statistically. It is impossible to support the claim that $\mu_1 - \mu_2 = 0$ or that $\mu_1 - \mu_2$ is equal to any specific value. However, other statistical approaches allow for claims that the differences between the groups are likely to be very small. Traditional NHST results fail to provide evidence of this sort. This suggests that researchers should look to alternative procedures to draw claims about the likely value of a parameter. The sections that follow contrast typical presentation that does not provide compelling support for null results to reporting and interpretation of confidence intervals that do allow for claims about null results.

APA Required Presentation

Following from Wilkerson and the Task Force on Statistical Inference (1999) and adopted in the current version of the APA publication manual (American Psychological Association, 2001), NHST results should be presented along with effect size measures. There are many estimates of effect sizes, the most popular being standardized difference estimates (e.g., Cohen's d ; Cohen, 1988) and proportion of variance statistics (e.g., eta-squared, omega-squared, and r^2). Here, I focus on Cohen's d and recommend its use as proportion of variance measures can be converted to d . A general strategy for interpreting d is to consider $d = 0.20$ as small, $d = 0.50$ as medium, and $d = 1.0$ as large (Cohen 1988).

Take the following fictitious examples wherein the researcher examined the effectiveness of an interactive computer-based tutorial for teaching about a particular statistical concept. Participants attended either a lecture on the topic (lecture group) or used a computer-based tutorial instead of attending the lecture (computer group). Following the completion of the lecture or tutorial, students took a 10-point quiz on the topic of interest. In this situation, the researcher wants to draw the conclusion that the tutorial is just as good as the lecture. Here a null result would be encouraging as it could suggest that the tutorial is comparable to a lecture. Here and below, we will consider two examples, both testing the null hypothesis that the difference between the tutorial and lecture groups is zero ($H_0 : \mu_1 - \mu_2 = 0$).

Example 1. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(108) = 0.2, p = .83, d = 0.04$.

Example 2. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(20) = 0.1, p = .93, d = 0.04$.

In both examples, the correct conclusion is to fail to reject H_0 . Additionally, the effect sizes are very small, suggesting we have no evidence that the groups differ in the population. However, based on this presentation, we cannot conclude much about the viability of the null hypothesis. The only appropriate statistical conclusion given the presentation above is that we have no evidence suggesting that the null is false. Here the presentation of the t , p , and effect size (d) does not transmit any information about whether H_0 is viable.

Recommendation One: Present Confidence Intervals around $\mu_1 - \mu_2$

The APA manual strongly recommends the use of confidence intervals (American Psychological Association, 2001).⁴ Examination of confidence limits around the differences between means improves interpretation considerably. This result will tell us what we could reasonably expect the differences to look like in the population. Estimates of confidence intervals around parameters are available through most major statistical analysis packages.

Loftus (1996) suggests that confidence intervals allow for the acceptance of a null hypothesis for all intents and purposes. Technically, we can not accept H_0 but confidence intervals can tell us whether differences between our groups would likely be meaningful or not. Put more simply, we may be able to say that the true deviation from H_0 is too small to worry about. This is a somewhat of a change in thinking from traditional NHST approaches wherein tests produce a simple dichotomy of reject/fail to reject. Below $\mu_1 - \mu_2$ represents the differences between the groups in the population.⁵

Example 1 with CI's. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(108) = 0.2, p = .83, 95\% CI = -0.6 \leq \mu_1 - \mu_2 \leq 0.7$.

Example 2 with CI's. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(108) = 0.2, p = .83, 95\% CI = -1.9 \leq \mu_1 - \mu_2 \leq 2.0$.

Consideration of confidence intervals tells us infinitely more than the t-test and effect sizes. For the first example, the confidence interval tells us that the differences could be as much as 0.6 points favoring the lecture group or 0.7 points favoring the tutorial group. On a ten-point quiz, we can interpret these differences as being relatively small; suggesting any difference between the tutorial and lecture is relatively unimportant. This does not technically support the null hypothesis that states the differences are zero. Rather, it gives us some idea how close to zero the actual population differences are. From this we can determine if the differences are large enough to worry about.

In Example 2, the differences could be as large as 1.9 points favoring the lecture group or 2.0 points favoring the tutorial group. On a ten-point quiz, two points would be a very large difference. This result provides little support for the null hypothesis, as large differences could reasonably exist in either direction in the population. There simply is not enough precision in the confidence interval to draw meaningful conclusions about the null hypothesis. Though both examples produce the same mean values and roughly the same t and p statistics, confidence intervals clarify the relationships considerably.

⁴ Unfortunately, the American Psychological Association does not require presentation of confidence intervals. Additionally, the most recent version of the publication manual does not indicate how to present intervals nor does the sample manuscript contain examples.

⁵ Note that the confidence intervals also transmit information regarding whether the null hypothesis should be rejected or retained. If the hypothesized difference between the means (0) falls outside of the confidence limits, we can reject the null hypothesis.

Recommendation Two: Present Confidence Intervals for Effect Sizes

Thompson (2002) takes recommendations regarding confidence interval presentation a step further and suggests presentation of confidence intervals for effect sizes. Interval estimates around effect sizes require specialized software or advanced programming of common statistical packages.⁷ Though based on concepts not traditionally covered in psychological statistics courses such as non-central distributions, the intervals are easy to interpret. The software recommended at the end of this paper make it easy to produce confidence intervals for effect sizes. For more information on these intervals see Smithson (2002).

Though confidence intervals around effect sizes convey some information similar to confidence intervals around the differences between means, they are superior as they provide a standardized metric allowing for comparisons across studies. Again, conclusions based on confidence limits do not focus on whether H_0 is viable; rather they represent how large a deviation from H_0 we can reasonably expect. Below, Δ represents the population effect size.

Example 1: Adding the CI around the effect size. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(108) = 0.2, p = .83, 95\% CI = -0.6 \leq \mu_1 - \mu_2 \leq 0.7, d = 0.04, 95\% CI = -0.33 \leq \Delta \leq 0.42$.

Example 2: Adding the CI around the effect size. The tutorial group ($M = 6.8$) and the lecture group ($M = 6.7$) did not differ, $t(108) = 0.2, p = .83, 95\% CI = -1.9 \leq \mu_1 - \mu_2 \leq 2.0, d = 0.04, 95\% CI = -0.80 \leq \Delta \leq 0.88$.

In both cases, the confidence interval around the effect size tells us a great deal about our results. Example 1 suggests that it is unlikely that the population effect size would be larger than $\Delta = 0.33$ favoring the lecture group or $\Delta = 0.42$ favoring the tutorial group. This means it is unlikely that any differences between the groups would constitute a medium-sized effect. This result does not necessarily support the null hypothesis but it does provide an indication of how big or small the effect might be in the population. Example 2 estimates that population effect size could be as large as $\Delta = -0.80$ favoring the lecture group or $\Delta = 0.88$ favoring the tutorial group. This result provides little evidence as to the viability of the null hypothesis, suggesting the population could take the form of a medium effect favoring the tutorial group, a medium effect favoring the lecture group, or any result somewhere in between.

⁷ Two great resources are available for calculating these confidence intervals. Exploratory Software for Confidence Intervals provides a demonstration version at <http://www.psylatrobe.edu.au/esci> and runs under MS Excel. SPSS syntax for calculations are available from <http://www.anu.edu.au/psychology/staff/mike/CIstuff/CI.html>.

Conclusions

Null hypothesis significance tests do not provide estimates of the actual value of a parameter; only what the parameter is unlikely to be. Therefore, presentation of results based on NHST procedures is inadequate to provide support for a null hypothesis. However, consideration of confidence limits does provide evidence as to how much we could reasonably expect the value of our parameter to deviate from the value specified in the null hypothesis. Thus, confidence intervals can support the viability of H_0 . Though it is not possible to claim that H_0 is true, we can sometimes conclude that our population value may be reasonably close to the value in the null hypothesis.

Perhaps the most important point of this paper is that we cannot draw conclusions supporting the null hypothesis without information such as confidence intervals around parameters and effect sizes. In reporting null results, confidence intervals around parameters and effect sizes are valuable tools. More importantly, these values allow greater precision in evaluation of results.

References

- * Indicates reference is highly recommended for further reading on the topic.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- * Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Harris, R. J. (1997). Reforming significance tests via three-valued logic. In L. L. Harris, S. A. Muliak, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Lawrence Erlbaum Associates.
- Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, 56, 16-26.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209.
- * Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161-171.
- Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of data. *Stevens' Handbook of Experimental Psychology*, Third Edition, Vol 4. (pp. 339-390). New York: John Wiley and Sons.
- * Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In L. L. Harris, S. A. Muliak, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 287-318). Mahwah, NJ: Lawrence Erlbaum Associates.
- * Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 2, 115-129.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.
- * Smithson, M. J. (2002). *Confidence intervals*. Thousand Oaks, CA: Sage.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31 (3), 24-31.
- Wilkinson, L. & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and Explanations. *American Psychologist*, 54, 594-604.

Send Manuscript Correspondence to

Dr. Chris Aberson, Ph.D.
 Department of Psychology
 Humboldt State University
 clab18@humboldt.edu

Received: November 8, 2002
Accepted: November 28, 2002