An Argument Framework for the Application of Null Hypothesis Statistical Testing in Support of Research

Steven D. LeMire
University of North Dakota

Journal of Statistics Education Volume 18, Number 2 (2010),
www.amstat.org/publications/jse/v18n2/lemire.pdf

Key Words: Null hypothesis; Statistical testing; Type I error rate; Type II error rate; *P* value; Argument; Framework; Teaching statistics.

Abstract

This paper proposes an argument framework for the teaching of null hypothesis statistical testing and its application in support of research. Elements of the Toulmin (1958) model of argument are used to illustrate the use of *p* values and Type I and Type II error rates in support of claims about statistical parameters and subject matter research constructs. By viewing the application of null hypothesis statistical testing within this framework, the language and intent of statistical support for research can be more precisely understood and taught.

1. Introduction

Null hypothesis statistical testing (NHST) is widely used in research and yet simultaneously criticized (Nickerson, 2000). Alternative methods for drawing statistical inference such as confidence intervals (CIs) and replication probabilities have been proposed and have also been criticized (Belia, Fidler, Williams, & Cumming, 2005; Miller, 2009; Wagenmakers & Grunwald, 2006). To evaluate the efficacy of various inferential methods in support of research, a better illustration of argument framework and how statistical evidence can be used to support claims about subject matter constructs is needed.

The intent of this article is to present a null hypothesis statistical testing argument framework (NHSTAF) which is informed by Toulmin (1958) argument elements that can better illustrate the support that NHST offers research construct claims. The Toulmin argument structure can be used

to identify the explicit purpose of the elements of NHST in relationship to arguments about subject matter constructs. This work is motivated by calls from authors for improving the education and understanding of the application of NHST. As an example of these calls, Nickerson (2000) has argued that a minimal goal should be to better understand NHST's application to experimental psychology. The American Psychological Association's Task Force on Statistical Inference called for elucidation of controversial statistical issues including NHST (Wilkinson & the Task Force on Statistical Inference, 1999). Cortina and Dunlap (1997) state, "The abuses of NHST have come about largely because of a lack of judgment or education with respect to those using the procedure. The cure lies in improving education and, consequently, judgment, not in abolishing the method" (p. 171). By presenting an explicit argument framework that address both the parameter and the subject matter research construct arguments and their relationship, educators will have an improved mechanism to teach the application of NHST.

The NHSTAF presented in this work illustrates how NHST supports claims about the subject matter construct as called for by Nickerson (2000). The demands on both the instructor and the students in teaching and learning the NHST procedure as it relates to claims about parameters can often leave little time for further conceptual development of NHST's support for subject matter research constructs. Because of this, student knowledge of NHST may not be contextualized into the broader scientific goals of providing support for real world research problems. This work presents a framework which can be adapted to instructional content to expand student contextualization and understanding of research claims supported by NHST.

NHST allows researchers to make decisions about statistical parameters and to quantify the chance that the decisions will be wrong. NHST takes advantage of our understanding of the nature of probability distributions of measured variables. Based on this understanding, researchers can construct inferential arguments about the parameters of distributions of observed scales and measures that are then used in arguments drawing conclusions about subject matter research constructs. Following Toulmin's (1958) precepts, in the statistical argument, the claim about rejecting the null hypothesis or failing to reject the null hypothesis is qualified by Type I or Type II error rates. Part of the validity of the inference of NHST is related to the correct quantification of these error rates. If used correctly, NHST can help guide researchers past chance variation to better optimize the research endeavor. In effect, a NHST puts up a test of chance that a study must pass. If passed, research, discussion, and effort may flow in the direction that the research illuminates. NHST is a small but important part of the entire research effort.

NHST has the potential to contribute positively to the advancement of research, yet a great deal has been written on its misuse in research. Many authors (e.g., Berger & Sellke, 1987; Carver, 1993; Cohen, 1990, 1994; Meehl, 1967; Nakagawa & Cuthill, 2007; Rozeboom, 1960; Thompson, 2004) have been critical of the application of NHST. This criticism also spans a wide area of research fields including: wildlife management (Johnson, 1999), physical therapy (Campo & Lichtman, 2008), communication research (Levine, Weber, Hullett, Park & Lindsey, 2008), and ecology (Fidler, Cumming, Burgman, & Thomason, 2004). The critics of NHST suggest that its use/misuse in research could actually impede the advancement of science. Hubbard and Ryan (2000), for example, would like to see NHST eliminated from curricula. They recommend that "graduate education programs must be modified to substantially reduce, if not

eradicate, the attention paid to SST [statistical significance testing]" (p. 677). Howard, Maxwell, and Fleming (2000) wrote, "Although the problems with NHST have been well documented, their use still dominates the research literature in psychology" (p. 315).

A great deal of the criticism concerning these problems is related to the misapplication and misunderstanding of NHST and, specifically, the use of *p* values by researchers, editors, and consumers of research. Nickerson (2000) noted, "Some point out that many of the criticisms of NHST are not so much criticisms of NHST *per se* but criticisms of some of its users and misuses. It is not the fault of the process, they contend, if some of its users misunderstand it, expect more from it than it promises to deliver, or apply it inappropriately" (p. 274). Nickerson discusses some of these misunderstandings, such as the belief that a small value of *p* means a treatment effect of large magnitude and that statistical significance means theoretical or practical significance. Nickerson also wrote that he believes that a key problem with the application of *p* values is a conceptual misunderstanding of the conditional nature of the *p* value.

Because of these types of concerns, confidence intervals (CIs) have been suggested as a replacement for NHST (APA, 2001). However, the advantage of CIs as a replacement for NHST with regard to correctly being understood by researchers is not clear. For example, Belia et al. (2005) suggest that a large proportion of published researchers do not understand how to apply CIs. They cite confusion between standard error bars and CIs and lack of understanding of the relationship between experimental design and CIs as examples, among others, of the difficulties researchers experience in applying CIs. They found these problems across disciplines and for both early and late career researchers.

In an attempt to better understand the application of CIs in research, Cumming and Finch (2005) offer some heuristic rules of thumb that are intended to be pragmatic guidelines for interpreting CIs. They say, for example, that "we should not make a big issue of whether a CI just includes or excludes a value of interest" (p. 177). Yet, a researcher still needs to interpret results and make decisions about the outcome of an experiment.

In order to diminish the role that NHST plays in research, both the American Journal of Public Health (AJPH) and Epidemiology (Epid) adopted policies whereby NHST was discouraged in submitted manuscripts. In these cases Fidler, Thomason, Cumming, Finch, and Leeman (2004) reported the following:

> In both journals, however, when CIs were reported, they were rarely used to interpret results or comment on precision. This rather ominous finding holds even for the most recent years we surveyed. In addition, in many AJPH articles in which NHST and *p* values were not explicitly reported, there was evidence, or at least clear hints, that interpretation was based on unreported NHST. (p. 123)

A major problem with the use of CIs to make research decisions is that CIs, by themselves, are not a decision-based mechanism like NHST. CIs are a mechanism providing interval estimates and can be useful when used with and without NHST. In fact, CIs are sometimes used as a surrogate for NHST, but only if one uses them to draw conclusions regarding parameters in the same way that one uses NHST. By themselves, CIs lack the argument structure of NHST. Because of this, one can expect that researchers' attempts to replace NHST with CIs will continue to be problematic.

In an effort to reduce the roll of NHST, the editors of the *Journal of Psychological Science* encourage authors to use p_{rep} instead of NHST. Killeen (2005) claims that "The p_{rep} statistic provides a graded measure of replicability that authorizes positive statements about results: 'This effect will replicate 100(p_{rep})% of the time' conveys useful information, whatever the value of p_{rep}" (p. 349). Killeen defines p_{rep} as a function of the p value as shown in equation 1.

$$p_{rep} \approx \left[1 + \left(\frac{p}{1-p}\right)^{2/3}\right]^{-1} \qquad (1)$$

Then, the p values of .05, .01, and .001 correspond to p_{rep} values .88, .95, and .99. Killeen calls these p_{rep} values "increments that are clear, interpretable, and manifestly important to a practicing scientist" (p. 349). Killeen further stated that "once p_{rep} is determined, calculation of traditional significance is a step backward" (p. 349).

In an article in *Psychological Science*, Brescoll and Uhlmann (2008) report on a study about gender and emotion, applying the recommendations by Killeen (2005). In their results they report on a two-by-three ANOVA investigating salary in relation to gender and emotion by stating "ANOVA revealed a significant interaction between the target's gender and expression, $F(2,112)=6.90$, $p_{rep}=.986$" (p. 273). A reasonable question is how did the authors conclude that the analysis was "significant" and what do they mean by the term? And, if the authors mapped the p_{rep} back onto the p value scale to make the conclusion for significance, what was their qualifier—Type I error rate—for their claim? It would seem that, like CIs, p_{rep} does not offer the necessary support for a complete argument in support of subject matter research claims.

Like NHST and CI, p_{rep} has been criticized as a measure of research argument support. Miller (2009) argues that although replications probabilities like p_{rep} are possible to estimate, they are almost never precise enough to be of use. This criticism of p_{rep} is supported by Iverson and Lee (2009), who state that p_{rep} misestimates the true probability of replication. Iverson and Lee conclude that p_{rep} is not a useful statistic for psychological science.

In order to better clarify the role of NHST in support of research, one should draw a clear distinction between the two main types of inference that are normally drawn in a great deal of research when NHST is used. These inferences regard conclusions about a statistical parameter or parameters and about subject matter research constructs. The relationship between these two inferences is often not clearly understood. Nickerson (2000) describes the confusion between epistemic (knowledge of subject matter construct) and statistical hypotheses. This confusion can lead to a misunderstanding of the use of NHST in support of research. Nickerson wrote, "The root of many of these problems [with NHST] is either confusion between epistemic and statistical hypotheses or a focus on the latter type to the neglect of the former" (p. 286). Bolles (1962) also describes these two inferences as different parts of the research process. Bolles identifies the statistical and the epistemic inferences as two different things. Meehl (1997) argued that researchers should distinguish statistical from epistemic questions. His distinction related to drawing inferences from a statistic to a parameter and then from an inferred parameter to a "substantive theory"(Meehl, p. 421). He urged that the distinction between these different kinds

of inference should be emphasized in statistics textbooks. The common point that these authors have emphasized is that researchers should pay close attention to the two types of inference that are used to construct arguments about distributional parameters and about subject matter constructs that are associated with research theories.

A better understanding of NHST is needed. Nickerson (2000) argues that NHST is the most widely used method of analysis in psychological experiments and if NHST is really as misunderstood as many researchers claim, it is an embarrassment. He recommends that we should seek a better understanding of the assumptions necessary to justify conclusions drawn from it. Stephens, Buskirk, Hayward, and Martinez Del Rio (2005) also stress that a better understanding of NHST is required. They state that "users of null-hypothesis tests, in particular, must greatly improve standards of reporting and interpretation" (p. 4).

This work attempts to clarify the use of NHST in support of research by using argument elements by Toulmin (1958). By doing so, the relationship between the inference about the parameter/s and the inference about the subject matter research construct can be better taught. In what follows, the Toulmin structure of argument elements will be introduced, and a research example will be discussed using the NHSTAF to draw an inference about a statistical parameter, the results of which will then be used to support an inference about a subject matter research construct.

2. Structure of Argument

There are many possible definitions of what constitutes an argument. Walton (1996) defines an argument as a sequence of propositions and as a claim for a conclusion justified by a premise that gives it support. For the purpose of illustrating the application of statistical support using NHST in research, four components of the Toulmin (1958) argument model will be used. Further complexities and arrangements of the Toulmin model elements are discussed by Miller, Nilsen, and Bettinghaus (1966).

For this work, we will be constructing an example argument about a statistical parameter that will be used in an argument about a psychological construct. The same argument structure could apply across a wide area of fields where NHST is applied. A statistical parameter is a characteristic of a distribution; in this example we will be using the mean of difference scores. The subject matter construct as discussed here is the idea of psychological depression. The elements of Toulmin used to construct these two arguments consist of data/evidence, warrant, claim, and qualifier. The data/evidence, warrant, and qualifier all support the claim.

The claim is the central part of the argument. A claim is the product of an assertion or conclusion put forth by the researcher (Toulmin, 1958). In NHST, a claim about a statistical parameter could be that the null hypothesis for a population mean equaling zero is false. In the case of the psychological argument dealing with the construct of depression, it could be that depression status decreased. Within a research context, a claim should be based on some form of facts which are called data/evidence.

In order to back up a claim, a researcher should be able to point to specific data/evidence. Data/evidence allows researchers to justify their claim. The data/evidence component includes explicit facts and other information that can be shared and can be used to support the claim. In the statistical argument about a parameter of a distribution, one might point to scores from study participants off a psychological scale. In the argument about a psychological construct, one could point to evidence of intervention and a statistical argument. Once you have the data/evidence and the claim, a question of interest is how the researcher justifies the claim from the data/evidence. The process that entitles a researcher to logically transition from the data to the claim is called the warrant (Toulmin, 1958).

There is a direct relationship between the data/evidence component and the claim that is logically linked or bridged by the warrant, which is the element that entitles the researcher to make claims from the data/evidence (Toulmin, 1958). While the data/evidence component is appealed to explicitly, the warrant can be implicit in the argument. Toulmin states that "the warrant is, in a sense, incidental and explanatory, its task being simply to register explicitly the legitimacy of the steps involved and to refer it back to the larger class of steps whose legitimacy is being presupposed" (p. 100). In the case of a dependent samples *t* test for a statistical argument, the claim may be to reject the null. The data could be difference scores from a depression scale. Key factors of the warrant that legitimizes the step from data to claim about the parameter are the statistical assumptions of the test. In the case of the dependent sample *t* test these assumptions are that the observed difference scores are independent and sampled from a normal population. It would be unlikely for a researcher presenting the results of a *t* test to add a paragraph justifying these assumptions. This is because it would be an implicit assumption that the researcher had prior evidence that the test would meet these assumptions. In practice one might use diagnostic tools, such as normal probability plots, to verify that these assumptions are not inconsistent with the data.

In the case of the argument about the construct, the claim may be that depression status decreased and part of the data/evidence could be the statistical argument where the null hypothesis was rejected. Part of the warrant that would bridge the data/evidence to the claim would be the assumption of construct validity of the scale used to assess depression status. Toulmin (1958) further describes the warrant as, "Our task is no longer to strengthen the ground on which our argument is constructed, but is rather to show that, taking these data as a starting point, the step to the original claim or conclusion is an appropriate and legitimate one" (p. 98). The warrant can be supported by other arguments that are not the central point of the current argument with warrant backing (Toulmin).

The backing of a warrant should give the consumer of the argument assurances as to its authority, so the step from the data/evidence to the claim is believed to be legitimate (Toulmin, 1958). Warrant backings are subarguments that are assumed to be valid. A warrant backing for the test of a parameter could be an argument for the appropriateness of the type of test. For example, this could be the argument for conducting a dependent sample *t* test instead of an independent sample test for a given case. For the subject matter construct argument, a researcher could present reliability estimates of other studies with similar participants for which the scale was used to give backing for the warrant establishing the appropriateness of the scale in measuring the proposed construct for the current population. Since scale reliability is a

prerequisite for the warrant of construct validity, this would be a legitimate form of warrant backing. With data/evidence, warrant, and claim we have the major components to strengthen an argument. The last component that we will need for this discussion is a mechanism to weaken an argument.

The qualifier restricts or limits the claim with what Toulmin (1958) calls conditions of exception or rebuttal. These are circumstances in which the claim would not be true or would be weakened through failures of circumstances of an implicit warrant, for example. The validity of inferential statistics conducted with NHST relies on the correct quantification of the chance the claim about the parameter/s is incorrect. For NHST, this qualification of claims is accomplished by the Type I and II error rates. The unit for error rate control is at the level of the claim about the subject matter construct. It can be argued that much of the misunderstanding and hence misuse of NHST, and consequentially its criticism, has its origin in a confusion of what is data/evidence, warrant, and qualifier and in which argument they are to be applied—parameter or subject matter construct.

3. Example of a Research Question Supported by NHST

Toulmin's (1958) argument structure will now be used in an example to illustrate the relationship between the four components of argument and the elements of NHST used to support conclusions in research; in particular, the example will illustrate the two arguments used to draw inferences about a statistical parameter and about a psychological construct. It is recognized that there are stronger designs that could be used in planning this type of study. In this study, for example, a randomized pre-post design looking at the interaction effect might have been more appropriate. However, the intent of this simple design is to illustrate how NHST can be used to support claims about subject matter constructs.

If a clinical psychologist wanted to study the efficacy of individual therapy on the depression status of new mothers, she could have mothers who previously have been diagnosed as depressed take the Beck Depression Inventory (BDI) (Beck, Rial, & Rickets, 1974) prior to the individual therapy (pre-assessment) and after the individual therapy (post-assessment). The researcher would like to make a claim regarding the therapy intervention interval in terms of the psychological construct of depression status as demonstrated by changes in the raw BDI scores from pre- to post-assessment.

There are two separate types of claims, and hence inferences, that the therapist will make in this study. The first claim will be about a statistical parameter that describes the quantitative scores assessed by the BDI scale. The second claim will be about the psychological construct of depression. The claim about the parameter in this example deals with mean population differences of the total BDI score. The claim about the psychological construct relates to the subject matter construct of depression status. These two claims are often combined in our thinking but require two separate, though related, arguments, with the argument about the statistical parameter supporting the argument about the psychological construct.

The validity of both of these arguments will be affected by the design employed for the study. As discussed above, a design with between and within factors might support stronger arguments for

both the statistics and the subject matter construct. This is because study design can affect plausibility of test assumptions. Study design is also especially important in relationship to claims of correlations (study) or causality (experiment) of an effect related to the subject matter construct.

The experimental unit for this example study is the individual mother who received the individual therapy and took the BDI independently of other mothers in the intervention. The population difference scores will be assumed here to be normally distributed and independent. The normality assumption relies on the premises that the construct of depression is continuous and at least an interval level variable. When the individual items on the BDI are summed for each individual, the population distribution of the scores is plausibly normal. The statistical hypothesis for the inference about the parameter is shown in Equation 2 (Ott, 1988).

$$H_O : \mu_d \geq 0$$
$$H_1 : \mu_d < 0 \qquad (2)$$

In this case, that inference will be a claim about the parameter μ_d being less than zero. The claim about the parameter will use a statistic $\bar{d}$ calculated from the data to draw inference about the population parameter μ_d. For the statistical test, the calculated statistics of the mean and the standard error of the difference scores are obtained, and the test statistic shown in Equation 3 is calculated. In this example, the expectation of $\bar{d}$, E(d-bar$_0$), is zero under the null hypothesis—this will not always be the case (Ott, 1988). In equation (2) μ_d is ≥ 0, not 0. To calculate type I errors one must restrict to point null hypothesis and it is customary to pick the value closest to the alternative. In this case this value would be zero and one would act as though the null hypothesis was $\mu_d = 0$.

$$t = \frac{\bar{d} - \mathrm{E}\,(\mathrm{d}\text{ - bar}_0)}{SE_d} \qquad (3)$$

Using this observed *t* value and the number of degrees of freedom associated with the standard error of the difference scores, a theoretical *t* distribution is referenced. From this theoretical distribution, a *p* value is computed under the assumption of the null hypothesis being true. The *p* value in this case is the probability of observing this value of *t* or smaller, given that the null hypothesis is true.

The *p* value is obtained from a theoretical distribution. This can be done because the population distribution is assumed to be normal. This assumption about the characteristics of the distribution will be part of the warrant of the argument dealing with the inference about the parameter. It is because of this assumption that we can calculate the chance probability of observing a given *t* value or smaller under the assumption that the null hypothesis is true. We reject the null hypothesis if the observed *t* value is not within the fail-to-reject region. The level at which we choose, prior to the study, the *t* statistic to be statistically significant will determine our

comparison-wise Type I error rate (α_{CW}), which will be part of the qualifier for the claim of rejection of the null hypothesis.

The therapist in this example hopes to demonstrate that the depression status of the new mothers decreased after the individual therapy interval. In order to have a complete set of arguments, she needs to calculate the necessary sample size for the fail to reject the null hypothesis possibility. She uses Equation 4 (Fisher & Van Belle, 1993) to approximate the sample size.

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\left(\frac{\bar{d}_1 - \text{E(d - bar}_0)}{\sigma_d}\right)^2} \quad (4)$$

Where $Z_{1-\alpha}$ and $Z_{1-\beta}$ are the (1-α) and (1-β) percentiles of the normal distribution, β is the desired Type II error rate; $\bar{d}_1$ and E(d-bar$_0$) are the expected values of $\bar{d}$ under a particular alternative hypothesis and null hypotheses respectively; and σ_d is the standard deviation of the population of difference scores.

The $Z_{1-\alpha}$, $Z_{1-\beta}$, $\bar{d}_1$ and E(d-bar$_0$) are all specified by the researcher based on knowledge of the requirements for the claim about the construct of depression status. The σ_d, however, has to be estimated. Cook and Campbell (1979) recommend the following process: "first, obtain agreement as to the magnitude of desired impact, and second, find acceptable variance estimates" (p. 40). The validity of the estimate of σ_d is part of the warrant for the claim about the statistical parameter when Type II error rate is used as a qualifier. Let's assume that the therapist felt that she would like to be able to find a statistically significant decrease in BDI scores (reject the null hypothesis of zero or positive difference) if the true population mean decrease in raw scale units was 5 points or more. She decided on 5 points or more because she felt that anything less would not be clinically significant, as discussed by Kazdin (1999) and Thompson (2002). For this study she references a previous study that had 54 participants, with a standard deviation of the difference scores of 12.6 on the BDI. Since she wants to be able to find an average decrease of 5 raw scale unit score points or more, her standardized effect size estimate is then 5/12.6 = 0.4. If she had wanted her standardized effect size to be 0.5 based on Cohen (1988), she then would expect the raw scale score decrease of 6.3 points or more on the BDI.

Cohen (1988) presented the standardized effect size as a simple way to estimate small, medium, and large effects. This standardized effect size is a mathematical convenience for thinking about effect size and for meta-analytic studies. It has many applications and can be a valuable tool. However, in the example used here, for support of the statistical claim and clarity of the statistical argument, the standardized effect size, stated without an actual estimate of standard deviation, can leave the researcher with an incomplete statistical argument and therefore incomplete support for the claim about the subject matter construct.

In this example, the qualifier for the statistical claim for the alternative hypothesis, when one fails to reject the null hypothesis based on a specified Type I error rate, is the Type II error rate. The Type II error rate is conditional on the alternative hypothesis that is defined in the a priori sample size calculation. This alternative hypothesis deals with the specified population mean difference in this example. It is not in terms of the standardized effect size. Therefore, the conditional element of the quantitative qualification of the claim is based on the specified mean difference in raw scale units from the sample size calculation. To have this specified effect size combined with the estimated population standard deviation confounds an estimate that does not have to be representative (the raw scale unit effect size) of the reality of the actual population with one that does have to be representative (the standard deviation). Since the accuracy of the standard deviation estimate is part of the warrant, to have it mixed in and confounded with the raw scale unit effect size limits its evaluation. And since there would be no specific a priori statement of the raw unit effect size, there will not be sufficient qualifying support for the claim about the alternative hypothesis through the statement of Type II error rate.

It is interesting to note that the only relationship to the actual data and study that this statistical power argument has, based on a failure to reject the null hypothesis, is through the warrant for the estimate of the standard deviation. Because prospective raw scale unit effect sizes can be specified independently of what the actual population parameter difference might be in raw scale units, a study that has low Type I error rate and low Type II error rate (high power) will not be more or less likely to reject the null hypothesis based on the data from the true population/s. In effect, a statistical test that has .20 power will not be more or less likely to reject the null hypothesis than one that has .99 power. And the only way to evaluate the correctness of a claim of statistical power in the statistical argument in this example is through the warrant for the estimated standard deviation of the test in the inference about the parameter.

As a further example, what if the researcher had used a Cohen (1988) standardized effect size of 0.5 without stating an estimate for the standard deviation of the difference scores in this example? Let's assume that she does her sample size calculation based on the standardized effect size. In her proposal, she could state that she had a .80 probability of finding a decrease of one half of a standard deviation. What if someone were to ask her what the actual mean change was in raw scale units in the BDI that she could expect to find? This would seem like a fair question, since the BDI raw scale scores will be used to assess change in depression status for the new mothers. However, if she had just used the standardized effect size, she could not answer that question. If she had done what the researcher actually did in this example, she then would have an estimate for the standard deviation of the difference scores and hence could state what mean raw scale unit effect size she could expect to find if this difference existed.

It still can be useful to use Cohen's (1988) small, medium, and large effect sizes and talk in terms of standardized effect size. This researcher had specified a standardized effect size of 0.4. In this case, the researcher actually thought about the relationship between the scale of the instrument and the argument about the statistical parameter and how a difference in outcome on the scale actually mapped into the fail to reject claim and the research construct. This is not easy to do, because it requires knowledge of the relationship between the measurement scale and the subject matter construct. This vital part of research is often skipped with the invocation of Cohen's standardized effect size. The omission of prior statistical power analysis can limit the

potential impact of research. As Maxwell (2004) writes, "If psychology is to continue to develop a coherent and accurate body of scientific literature, it is imperative that further attention be given to the role of power in designing studies and interpreting results" (p. 161).

The researcher is now ready to determine the sample size for her study. She uses the estimate for the standard deviation of the BDI scores from previous depressed mothers to estimate the variability in the difference scores, σ_d . She then specifies the raw scale unit effect size of five, a Type II error rate of .2, and a Type I error rate of .05 to determine the sample size necessary for the study.

Once the sample size calculation has been completed, the researcher recruits the number of representative depressed new mothers necessary and begins the study. Within this design, the researcher has specified prior to conducting the study her Type I and Type II error rates. These error rates represent quantitative qualifications of claims for two potential separate arguments about the difference parameter of the population distribution of BDI scores.

The data are then collected, the statistical test is completed, and the null hypothesis is rejected or fails to be rejected. The Toulmin (1958) structure is first applied to the arguments examined in the case when the null hypothesis is rejected as shown in Figure 1. Then, the framework is examined in the case when the null hypothesis fails to be rejected as shown in Figure 2. Each of these figures shows an argument framework for the construct and a statistical parameter. In both figures, the claim about the subject matter construct is the central point of the research question. Because of this, the construct argument is shown at the top and the statistical argument is shown at the bottom. Once completed, the statistical argument is used to support the construct argument.

4. Arguments

4.1 Statistical Argument When the Null Hypothesis is Rejected

Assume that the null hypothesis is rejected, whereby the observed level of significance (p) was less than α_{CW} ($p<\alpha_{CW}$). The claim being made is that the population parameter of the mean of the difference scores is less than zero. This inference regards the population parameter μ_d . The data and evidence being used in the statistical inference about the population parameter are derived from the difference scores of the BDI measure of depression. The statistics for the mean and standard error of the difference scores were calculated. The test statistic was determined from these statistics and was used to calculate the p value under the assumption that the null hypothesis is true. The test statistic and the p value are the data and evidence part of the argument about the population parameter. The warrant is the bridge from the data and evidence to the claim. Part of the warrant in this example is the assumption that the population distribution of the difference scores is normal. An additional element of the warrant for the claim about the parameter is that the individual difference scores are independent. If either of these conditions does not hold, the bridge or link between the data/evidence and the claim about the parameter could not be substantiated for the dependent sample t test. The backing for these warrants could involve a discussion of how the mothers went through the therapy independently and how there

were no distributional problems expected with the data (i.e., ceiling or floor effects). The qualifier for the claim about the parameter is the statement of comparison-wise Type I error rate for this test. An illustration for the argument about the parameter when the null hypothesis is rejected is shown in Figure 1 (Statistical Argument Reject).

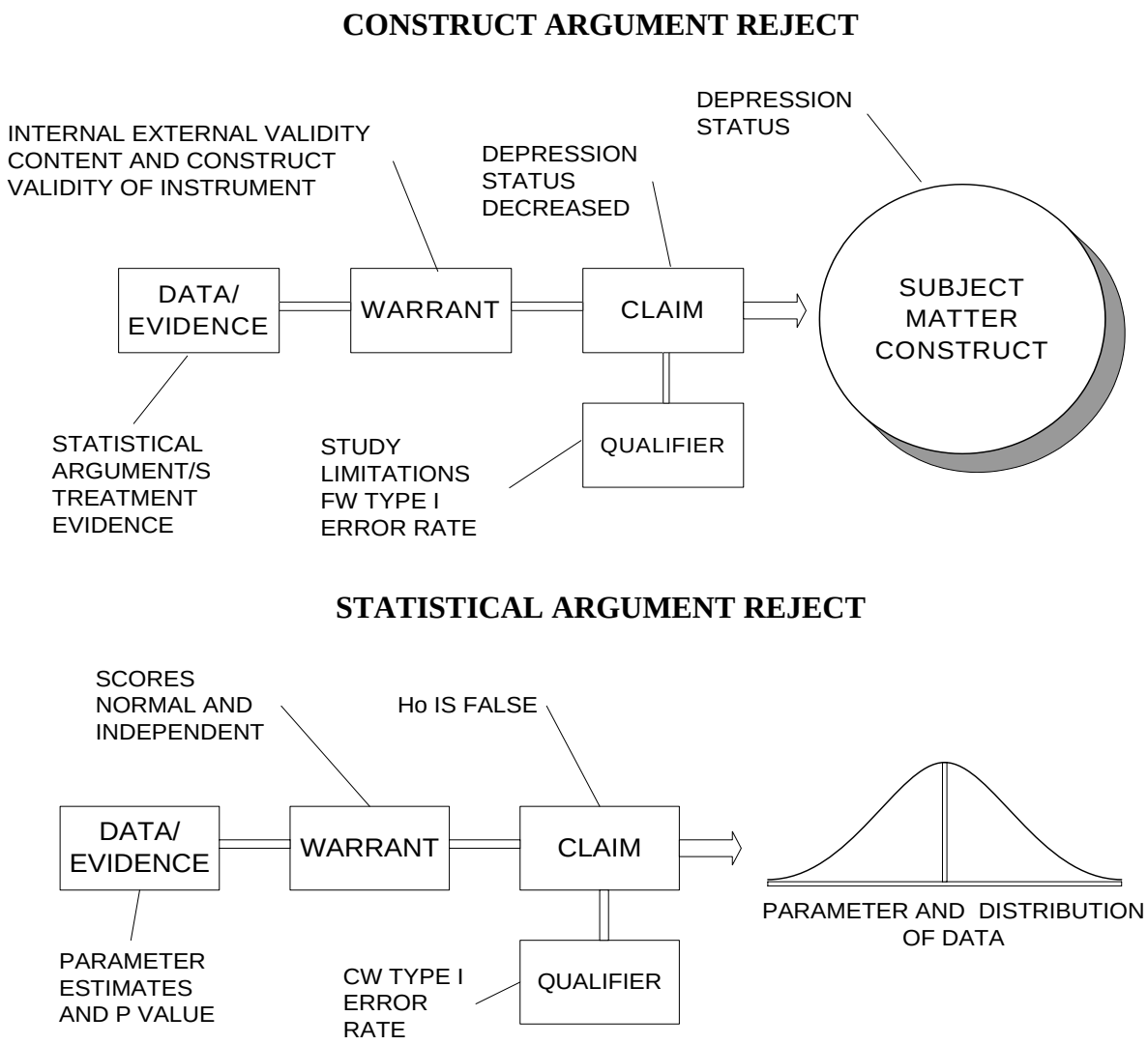


Figure 1. Toulmin model elements used in arguments when null hypothesis rejected. (CW: Comparisonwise, FW: Familywise)

When the null hypothesis is rejected, the statistical claim for this example is that the parameter (μ_d) is less than zero. The observed difference can be any value over the distributional range of the test statistic. The *p* value can be used to state explicitly the chance of observing any given range of the test statistic, conditional on the null hypothesis being true and valid distributional assumptions in the warrant. When statistical significance is claimed, it means that the probability of observing this value or one more extreme is sufficiently small that it will be concluded the null hypothesis is false. Sufficiently small is defined by the Type I error rate, which is the

qualifier for rejection of the null hypothesis. This claim is qualified with the a priori comparison-wise Type I error rate for this individual test.

4.2 Construct Argument When the Null Hypothesis is Rejected

To this point, an argument has been made concerning a parameter of a distribution. The therapist of the study wants to make a claim about the psychological construct of depression with regard to the individual therapy treatment. For this claim, part of the data and evidence for the inference about her psychological construct of depression is the argument about the parameter of the difference scores from the scale used for the evaluation of depression. Other elements of the data for the argument about the construct might be evidence that the new mothers actually received the therapy. Warrants given for the claim would be that the instruments had validity and reliability. Another warrant might be that these depressed mothers were a representative sample of depressed mothers in general. These warrants are often elements of what are considered to be matters of internal and external validity. Backing for the warrant of reliability might be a psychometric measure of reliability demonstrated on this population and this scale. The qualifier for the claim would be made up of a number of elements as well. One of the qualifiers for the claim in this example would be the family-wise Type I error rate controlled at the level of the claim about the subject matter construct. The claim about the subject matter construct is the unit for the Type I error rate control. In this case, only one statistical test was completed that could have led to a successful claim about the construct. Because of this, the comparison-wise Type I error rate equals the family-wise Type I error rate for this example. If more than one statistical test was conducted, and if any one of those tests could have allowed for a successful argument at the construct level, then the comparison-wise Type I error rate would not equal the family-wise Type I error rate. An illustration of the argument about the construct is shown in Figure 1, which is titled Construct Argument Reject.

The difference between the two arguments in Figure 1 is that one argument deals with the claim and inference about a parameter whereas the other argument deals with the construct of depression status for new mothers. The argument about the subject matter construct uses the argument about the parameter as data/evidence. Even if the instruments did not have content validity, the statistical argument about the population parameter could still be valid. This is because this argument was only concerned with the distribution of the difference scores obtained from the BDI scale. The argument about the psychological construct of depression status would then be incorrect, because the warrant which connects those scores to the claim about the construct of depression status would not have been valid if the instrument did not have validity.

In this case, the investigator would have observed some $\bar{d}$ from the data. She could claim that depression status decreased (t=observed t, df=degrees of freedom, $p<\alpha_{FW}$). Her best estimate of that change would be the raw unit $\bar{d}$ or the standardized effect size ($\bar{d}/S_d$). At this point, she could claim statistical significance, but being able to claim practical significance would depend on whether $\bar{d}$ or $\bar{d}/S_d$ was large enough to be practically or clinically important (Thompson, 2002). The rejection of the null hypothesis would be one piece of data/evidence to be used for the argument about the subject matter construct. The effect size and a confidence interval are additional pieces of data/evidence for support of the construct argument.

4.3 *p* Value's Use in Argument about Parameters

Since a great deal of the criticism of NHST is derived from the misapplication of the *p* value, it will be discussed further here. The *p* value is the probability of obtaining a test statistic as extreme or more extreme (in this example smaller) than observed, given that the null hypothesis is true. It can be used as data/evidence in the argument about the parameter. Given the data/evidence, a decision about the null hypothesis will need to be made. If a claim is made that the null hypothesis is false, the qualifier for this claim is not the *p* value but the a priori Type I error rate. This is the level, which was decided prior to looking at the outcome of the test, at which the null hypothesis would be rejected. The *p* value is part of the data or evidence for the statistical argument, but it is not part of the qualifier. If the researcher had made the subject matter construct claim that depression status decreased (*t*=observed *t*, df=degrees of freedom, *p*=.002) she would be placing data/evidence for the statistical argument into the argument about the subject matter claim where p=.002 has no interpretation for construct argument qualification or practical significance of a possible effect related to the subject matter construct.

Discussion of the presentation of results for what is data/evidence and what is qualifier in NHST is not a new one. Gigerenzer (2004) discusses Fisher and the Neyman and Pearson approaches. He says, "For Fisher, the exact level of significance (the *p* value) is a property of the data, that is, a relation between a body of data and a theory" (p. 593). The "theory" here is assumed to be the distributional assumption of the test statistic under the null hypothesis. Gigerenzer goes on to say that, "For Neyman and Pearson, α (Type I error rate) is a property of the test, not of the data" (p. 593). In the argument structure present here, it is clear that the Type I error rate is a qualifier and is not data/evidence, as the *p* value is.

By just presenting the *p* value, a researcher does not give the necessary qualifier for a claim. For example, an author might report a test for a parameter as statistically significant, $t(16) = 2.12$, $p = .025$. To this point, the author has only reported the *p* value or level of significance. The *p* value by itself, however, does not qualify the claim about the parameter. By just reporting the *p* value, the researcher fails to give the quantitative qualifier for the comparison-wise Type I error rate for the claim about the parameter; therefore, the argument is left incomplete (see Figure 1, Statistical Argument Reject). And, since Type I error needs to be controlled at the level of the argument about the subject matter construct, this type of presentation—one where comparisonwise (and for the argument about the subject matter construct, familywise) Type I error rate is not specified—implies that the researcher is not taking responsibility for the statistical testing being conducted.

Because the *p* value is only part of the overall data/evidence of an argument, the researcher who conducted the study is in the best position to make a claim about the parameter that supports the claim about the subject matter construct. This is because the researcher has the best understanding of the warrants and the warrant backing that are necessary to give credibility to the claim. And since Type I error rate is controlled at the level of the claim about the subject matter construct, strict control over the number of tests used for a family must be maintained to account for the actual Type I error rate reported in the qualification of the claim about the subject matter construct. It is unlikely that a consumer of this research will have this necessary insight in

order to make appropriate and complete claims about the parameter/s and the construct if *p* values are presented for statistical tests without Type I error rate qualification. Because of this, the approach of just presenting *p* values does not represent a complete argument structure by the researcher.

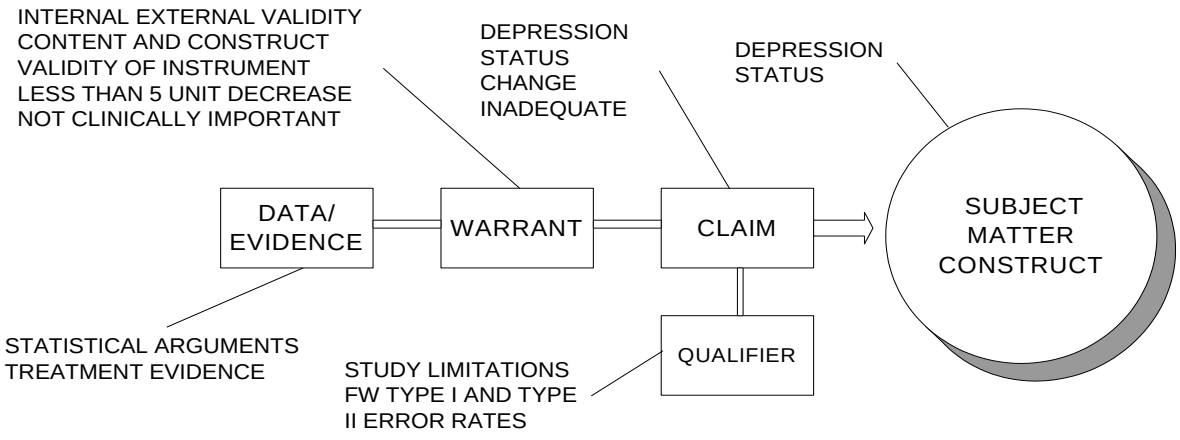
4.4 Statistical Argument When the Null Hypothesis is Not Rejected

When the null hypothesis was rejected, the claim about the parameter was that it was less than zero in this example, and this claim was qualified with a statement of Type I error rate. When one fails to reject the null hypothesis, a Type I error cannot be made. However, a Type II error can be made. Often the claim concerning Type II error rate (β) is replaced by a claim regarding statistical power, where power $=1-\beta$. By itself, failure to reject the null hypothesis does not mean that the researcher can make a successful argument that the parameter in this example is zero or greater than zero. Finkelstein and Levin (1990) wrote the following on this topic:

> Frequently a finding of non-significance is interpreted as support for the null hypothesis. This may be warranted if the power to detect important effects is high, for then the failure to detect such effects cannot easily be attributed to chance. But, if power is low, then the non-significant results cannot be taken as support of either hypothesis and no inference should be drawn from the failure to reject the null hypothesis. (p. 187)

An argument often made in the nonrejection context relates to claiming that if there was an important difference, the null hypothesis would have been rejected. In this case the researcher completed a sample size calculation whereby she specified an a priori Type II error rate and a raw scale unit effect size within the sample size calculation. The effect size was what she believed to be a clinically important raw scale unit difference in BDI scores. To carry out the sample size calculation, the researcher fixed the raw scale unit effect size and Type I and II error rates and used an estimate of the standard deviation of the difference scores for the sample size calculation.

CONSTRUCT ARGUMENT FAIL TO REJECT



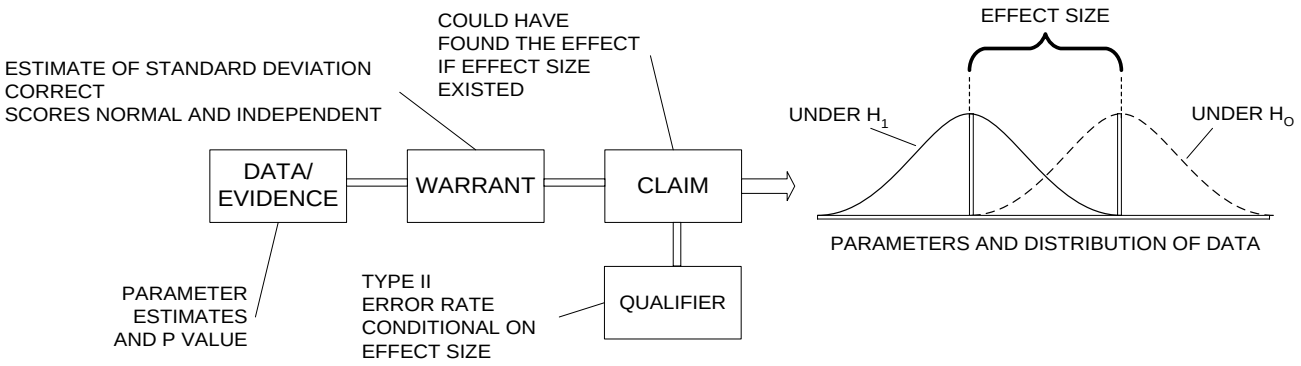


Figure 2. Toulmin model elements used in argument when null hypothesis failed to be rejected. (CW: Comparisonwise, FW: Familywise)

The statistical claim, when failing to reject the null hypothesis after doing the a priori sample size calculation, is that she could have rejected the null hypothesis if the prespecified raw scale unit effect size or larger existed. The argument about the statistical parameter when the null hypothesis is not rejected is illustrated in Figure 2 (Statistical Argument Fail to Reject). In this case the important effect size was a five point or more average decrease from the pre-assessment to the post-assessment. The qualifier for this claim is the Type II error rate. Type II error rate is conditional on the specific alternative to the null hypothesis being true. The alternative to the null hypothesis used to calculate the sample size is the researcher's a priori stated effect size in raw scale units of five points on the BDI. The evidence for this argument is the a priori sample size calculation with the estimated standard deviation. One of the warrants is that the actual estimate used for the standard deviation is correct. The assumptions that the data were normally distributed and independent are also part of the warrant for this example.

4.5 Construct Argument When the Null Hypothesis is Not Rejected

The argument about the construct when the null hypothesis is not rejected is illustrated in Figure 2 (Construct Argument Fail to Reject). In this example, the claim would be that the depression status change due to the intervention interval was inadequate. The researcher would qualify this claim with Type II error rate and a statement of what is meant by a meaningful effect size difference. This is because Type II error rate is conditional on the alternative hypothesis, and this is defined as the effect size in raw scale units in the a priori sample size calculation.

So in the case when the null hypothesis was not rejected, the argument about the psychological construct of depression status would center on the claim that depression status decrease was inadequate. The researcher could claim that she should have been able to find a difference (reject the null hypothesis) if the decrease was five points or more with a Type II error rate of β for the quantitative qualifier for this claim. Or as is often stated, she could say that she had (1- β) power to find a difference (reject the null hypothesis) if the population mean decrease was five points or more. So, she might present a statement that no clinically meaningful decrease in depression status was demonstrated for the individual therapy session interval (t=observed t, df=degrees of

freedom, $p > \alpha_{FW}$) with 80 percent power to reject the null hypothesis if an average decrease in raw scale units of five points or more existed. Here she qualifies her claim about the psychological construct with the Type I error rate (α_{FW}) and one minus the Type II error rate (1-β).

5. Discussion

When null hypothesis statistical testing is to be used to support an argument about a subject matter construct, all 16 elements of Figures 1 and 2 should be addressed. These figures illustrate that there are two distinct arguments dealing with a statistical parameter and a subject matter research construct. In the case of rejecting the null hypothesis, the claim about an individual parameter is qualified by the comparisonwise Type I error rate. The claim about a subject matter research construct is, in part, qualified by the familywise Type I error rate. The quantification of the chance that the statistical conclusion could be wrong, controlled at the level of the claim about the subject matter construct by the a priori specified familywise Type I error rate, is the necessary qualifier for the research claim when null hypothesis statistical testing is used. The *p* value is not the qualifier in either of these arguments.

The NHSTAF illustrated here can help clarify the use of NHST in the support of research as called for by Nickerson (2000). The framework explicitly illustrates the parallel argument structure that is conceptually necessary to apply NHST in support of subject matter research constructs. This framework addresses the need for distinguishing the statistical from the epistemic questions as called for by Meehl (1997).

An area where NHSTAF could have a large impact on the improvement of reporting standards for null hypothesis statistical testing is in relation to the use of the *p* value. This could happen through a better understanding that a statistical test/s supports a claim about a subject matter construct and that the researcher is responsible for Type I error rates qualification for the subject matter construct claims. An example of this misunderstanding is an author stating that a table of tests was significant at *p<.05, **p<.01, ***p<.001 which would imply a simultaneous qualifying Type I error rate of .05, .01, and .001. If the researcher would have given a test a *(p<.05) and rejected at the .05 Type I error rate but obtained a *p* value less than .01, it would be a misrepresentation of qualification to report the test as if the Type I error rate was .01.

An author would be expected to correctly qualify the Type I error rate for a null hypothesis statistical testing test/s and then for the claim about the subject matter construct. An editor might respond to the *p<.05, **p<.01, and ***p<.001 table with the question: what subject matter construct claim did these NHST tests support and what were your specific comparisonwise and familywise Type I error rates?

This framework also fills a pedagogical need for the teaching of the application of null hypothesis statistical testing in support of research. A major area often missing in student knowledge is the conceptualization of the unit of control for Type I and Type II error. By applying this structure in the teaching of null hypothesis statistical testing, instructors will have a

better framework to improve the standards of reporting and interpretation of null hypothesis statistical testing by students as called for by Stephens et al. (2005).

6. Example of Teaching Unit

There are obviously many possible approaches to teaching NHST. The argument framework presented within this work will not replace current instructional content used for teaching but can supplement and extend the logic of NHST for the student by connecting it to a broader framework of their argument knowledge. A goal should be to assist the student in seeing the logical argument framework of NHST. This would include understanding what is data/evidence, warrant, claim, and qualifier within the two argument structure, related to a claim about a statistical parameter/s supporting a claim about a subject matter research construct.

I give an example below of a short introductory teaching segment for non-statistics major graduate students. A major goal of this initial introductory instruction is to place the NHSTAF into the students' contextual knowledge of argument. It is my belief that a lot of the criticism that NHST faces is due to the misunderstanding of the purpose of *p* values. This includes the lack of understanding of the implementation of a qualifier.

To introduce the topic, I often start with the statement, "I have some land in a nearby county and I am thinking about buying a hog." I then ask the students what I am going to buy. Invariably, some students claim that I am going to buy a motorcycle (Harley-Davidson) and some claim that I am going to buy a farm animal. We then talk about the data and evidence that I provided them with my statement and how they came to their claims as to what I was going to buy. This leads to a discussion of their implicit warrants for their claims—which we later relate to statistical test assumptions. We then assess if anyone had said they *thought* I was going to buy one thing or the other or if they just said I was going to buy an item. For those students who said *I think* that I was going to buy a given item, we discuss how that limited the strength of their claim—which we later related to Type I error rate.

After the initial discussion of argument structure, I have the students construct an argument specific to their individual research fields using the Toulmin (1958) structure. Students normally have a sound understanding of the nature of data/evidence and claims but have less understanding about warrants and qualification of claims. After discussion of student arguments, we set up a research question and define a statistical hypothesis. For this exercise, the students are given Figure 1. Throughout this exercise, I continually relate the elements of the NHSTAF back to the students' example arguments.

To evaluate the impact this has on student perceptions after presenting this instruction, I obtained instructional review board approval to ask 24 nonstatistics major graduate students taking midlevel statistics the questions shown in Table 1. The students were given the options of: strongly agree, agree, slightly agree, slightly disagree, disagree, and strongly disagree for each question.

Table 1. Percentage of Some Form of Agreement of Initial Instructional Unit of Null Hypothesis Statistical Testing Argument Framework (slightly agree, agree, strongly agree for some form of agreement).

		% of Some Form of Agreement N=24
Q1.	Null hypothesis statistical testing is an important procedure for me.	83
Q2.	The Toulmin argument model helped me better understand the logic of null hypothesis statistical testing.	79
Q3.	I was able to connect my understanding of everyday argument with the logic of null hypothesis statistical testing.	79
Q4.	The exercise of first constructing an everyday argument and then using the same logic structure to complete a statistical argument helped me gain a better understanding of null hypothesis statistical testing.	92
Q5.	I will continue to use the Toulmin argument structure to learn the application of null hypothesis statistical testing.	83
Q6.	My better understanding of the Toulmin argument structure applied to null hypothesis testing will improve my confidence in my ability to use statistical testing to support my research.	79

It is interesting to note that only 83% of the students indicated that NHST was an important procedure for them. Many students take this course because it is required in their program. A majority of the students indicated a better understanding of the application of NHST based on this introductory instructional segment. The highest agreement came for the question which assessed connecting the logic of the NHSTAF to their real world understanding of argument. This is a goal of contextual learning where instructional content is presented in context of the individual student's prior knowledge (Sears, 2002). What we would like to do is bridge from the student's knowledge of NHST to their application of it through a strong comprehension of the logical framework—which is presented in this work. After this initial introduction instructional segment discussed here, I would refer back to the argument structure throughout the course and gradually expand on it to further include multiple comparisons and the fail to reject elements.

The goal of continually revisiting the argument framework throughout the course is to promote deep conceptual understanding of NHST in support of subject matter claims. This is the goal of Garfield and Everson (2009) when they urge statistics teachers to stress conceptual understanding rather than mere knowledge of procedures. It has been my experience that students interested in quantitative methods have greatly benefited from the implementation of the NHSTAF within my statistics courses. It has helped them understand the roll of NHST and help differentiate its contribution from other important information such as effect sizes, CIs, and CIs of effect sizes (Kelley, 2007; Thompson, 2007).

7. Conclusion

A metaphor for the statistical argument may be a flying buttress supporting a medieval cathedral. The cathedral is the argument about the subject matter construct built with the theoretical foundations of the specific science. Valid statistical arguments are part of the support for the argument about the subject matter construct. A statistical argument by itself, without a subject matter construct argument that is motivated by a substantive theory, would collapse without meaning, just as would a flying buttress by itself. The statistical tests are defined in relation to the subject matter construct argument's claim. Other data/evidence, like effect sizes and CIs, can and should be simultaneously used as support for an argument about a subject matter construct.

Finally, we seldom can offer proofs while doing research. What we can do is build arguments that support specific conclusions. NHST, CIs, and effect sizes can all support arguments about subject matter constructs. The logic of NHST, with its ability to support claims with quantifiable qualification criteria for both the reject and failure to reject cases at both the parameter (comparisonwise) and subject matter construct (familywise) levels, makes it a logical tool for a wide area of research fields. A researcher doing one statistical test in support of a claim about a subject matter construct should consider all 16 argument elements shown in Figures 1 and 2.

It has been over 16 years since Cohen (1994) stated that NHST had failed to support the advancement of his field's science. He also said that we should not look for a "magical alternative" to NHST because it does not exist (p. 1001). And today, NHST is still widely used and debated. What is more likely the problem is the way NHST is taught and assumed to be learned by students. The statistical education community needs to seek out better ways to connect our science with the correct limitations, knowledge, comprehension, and application of scientific users of NHST.

References

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

Beck, A. T., Rial, W. Y., & Rickets, K. (1974). Short form of depression inventory: Cross-validation. *Psychological Reports, 34*, 1184-1186.

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

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association, 82*, 112-139.

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11*, 639-645.

Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead?: Status conferral, gender, and expression of emotion in the workplace. *Psychological Science, 19,* 268-275.

Campo, M., & Lichtman, S. W. (2008). Interpretation of research in physical therapy: Limitations of null hypothesis significance testing. *Journal of Physical Therapy Education, 22*, 43-48.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 64*, 287-292.

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

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

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*(2), 161-175.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170-180.

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics, 33*, 615-630.

Fidler, F., Thomason, N., Cumming, G., 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.

Finkelstein, M. O., & Levin, B. (1990). *Statistics for lawyers*. New York, NY: Springer-Verlag.

Fisher, L. D., & Van Belle, G. (1993). *Biostatistics: A methodology for the health sciences.* New York NY: John Wiley & Sons.

Garfield, J., & Everson, M. (2009). Preparing teachers of statistics: A graduate course for future teachers. *Journal of Statistics Education, 17.*

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5*, 315-332.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681.

Iverson, G. J., & Lee, M. D. (2009). P_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review, 16,* 424-429.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*(3), 763-772.

Kazdin, A. E. (1999). The meaning and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67(3),* 332-339.

Kelly, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8).

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*(5), 345-353.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Massi Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34(3),* 171-187.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34(2),* 103-115.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-425). Mahwah, NJ: Erlbaum.

Miller, G. R., Nilsen, T. R., & Bettinghaus, E. P. (1966). *Perspectives on argumentation.* Chicago, IL: Scott, Foresman.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16(4),* 617-640.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence intervals and statistical significance: A practical guide for biologists. *Biological Reviews, 82(4),* 591-605.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Ott, R. L. (1988). *An introduction to statistical methods and data analysis* (3rd ed.). Boston, MA: Wadsworth.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*(5), 416-428.

Sears, S. (2002). *Contextual teaching and learning: A primer for effective instruction*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Martinez Del Rio, C. (2005). Information theory and hypothesis testing: A call for pluralism. *Journal of Applied Ecology, 42*, 4-12.

Thompson, B. (2002). "Statistical, Practical, and Clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Thompson, B. (2004). The "significance" crisis in psychology and education. *The Journal of Socio-Economics, 33*, 607-613.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423-432.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

Wagenmakers, E. J., & Grunwald, P. (2006). A bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17(7)*, 641-642.

Walton, D. (1996). *Argument structure: A pragmatic theory*. Toronto Canada: University of Toronto Press.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594 – 604.

Steven D. LeMire
University of North Dakota
mailto:steven.lemire@und.nodak.edu