Testing for Statistical and Practical Significance: A Suggested Technique Using a Randomization Test

John Fraas, Ashland University
Isadore Newman, University of Akron

Available at: https://works.bepress.com/john_fraas/15/
Running head: TESTING FOR STATISTICAL AND PRACTICAL SIGNIFICANCE

Testing for Statistical and Practical Significance:
A Suggested Technique Using a Randomization Test

John W. Fraas
Ashland University

Isadore Newman
The University of Akron

Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL

October, 2000
Abstract

A testing procedure is presented that uses a randomization test to statistically test non-nil null hypotheses which incorporate values deemed to be practically significant. This procedure stresses two important philosophical positions. First, the concepts of practical and statistical significance are both essential components of the evaluation process. Second, practitioners and researchers should view the process of establishing the level of practical significance not only as a statistical one, but also as one in which they give consideration to societal concerns and cost versus benefit comparisons.
Testing for Statistical and Practical Significance:

A Suggested Technique Using a Randomization Test

We believe that current research practices will be strengthened if researchers incorporate into their work the use of nil-null hypotheses that are based on effect sizes deemed important by researchers and practitioners. Thus, we are proposing that researchers consider using a testing procedure that incorporates three key elements. The first element is the use of a non-nil null hypothesis, i.e., a null hypothesis in which the test value is not zero but rather some value of importance or interest to the researchers. The second element is the determination of a practically significant level that is incorporated into the non-nil null hypothesis. The third element is the use of a randomization test to statistically test the non-nil null hypothesis.

This procedure stresses two important philosophical positions. First, the concepts of practical and statistical significance are both key components in the evaluation process. Second, it is important for practitioners and researchers to consider the process of establishing the level of practical significance not only as a statistical one, but also as one in which concerns more qualitative in nature are considered.

Null Hypothesis Statistical Testing.

As noted by Kirk (1996), “for almost 70 years, null hypothesis significance testing has been an integral part of the research enterprise in which behavioral and educational researchers engage. And for almost 70 years, null hypothesis significance testing has been surrounded by controversy” (p. 746). Berkson published an article in 1938 that provided one of the earliest challenges to the use of null hypothesis statistical testing. More recently, numerous authors have challenged the use of null hypothesis statistical testing (Carver, 1978, 1993; Cohen, 1990, 1994;

Thompson (1999a) stated: “A few scholars have called for the banning of statistical significance tests. However, the fact that many psychologists misinterpret statistical significance tests is not a reasonable warrant for banning these tests. Consequently, attention has now turned toward ways to improve practice” (p.169). Through our proposed testing procedure, it is our intention not to abandon statistical testing procedures but rather to improve them based on recommended changes discussed in the literature.

**Recommended Changes in Current Research Practices**

Various researchers have suggested ways to improve or supplement current research practices. We believe that three of the recommended changes are noteworthy. First, as suggested by numerous authors, including Cohen (1988, 1994), Huberty (1993), Robinson and Levin (1997), Shaver (1993), and Thompson (1996, 1999a, 1999b, 1999c, 2000), current research practice should incorporate the reporting and interpreting of the effect sizes, i.e., measures of practical significance. Second, as argued by Robinson and Levin (1997), the results of statistical hypothesis testing should be conducted and reported along with the effect sizes. Third, as recommended by Cohen (1994), researchers should use non-nil null hypothesis. Cohen stated: “Even null hypothesis testing complete with power analysis can be useful if we abandon the rejection of the point nil hypotheses” (p. 1002).
Effect sizes versus statistical significance. Currently, the reporting of effect sizes, which is considered as a method of addressing practical significance, is strongly supported by researchers. Such support can be found in recent suggested changes in research practices by various journals and a recommend change in the editorial policy of the American Psychological Association. As Thompson (1997), the editor of *Educational and Psychological Measurement*, stated: “As an editor, I do not reject articles reporting the results of statistical significance tests. However, I do expect to see effect sizes . . . reported and interpreted” (pp. 31-32). The Association for Assessment in Counseling (1990) stated the following guidelines for authors who publish in the journal *Measurement and Evaluation in Counseling and Development*: “7. Authors are strongly encouraged to provide readers with effect size estimates as well as statistical significance tests. . . . 8. Studies in which statistical significance is not achieved will still be seriously considered for publication” (p. 48). The change in editorial policy of the American Psychological Association is evident in its 1994 APA style manual which states: “You are encouraged to provide effect-size information” (APA, 1994, p. 18).

Importance of conducting statistical hypothesis testing. Less agreement, however, has been reached with respect to the importance of conducting statistical hypothesis testing when effect sizes are reported and interpreted. Shaver (1993) and Thompson (1999c) expressed the view that formal statistical hypothesis testing might be an optional companion to the reporting of effect sizes, i.e., practical significance. Shaver expressed the view that: “In short, studies should be published without tests of statistical significance, but not without effect sizes” (p.311). Thompson echoed a similar view: “My view is that statistical significance testing is in many
respects often merely irrelevant. I don’t object to statistical tests as long as effect sizes of some flavor are always reported” (p. 159).

Levin (1998), Levin and Robinson (2000), and Robinson and Levin (1997) took issue with the position taken by Shaver (1993) and Thompson (1999c). Robinson and Levin expressed the view that declarations of statistical significance should regularly precede deliberations of substantive significance. In light of this position, Robinson and Levin (1997) proposed a two-step data analysis process. In this two-step procedure the researchers would first determine whether the observed effect was statistically significant. Only if the observed effect was statistically significant would the researchers implement the second step in which they would assess the practical significance of the observed effect.

Reasons for the dearth of non-nil hypotheses in current research. Thompson (1999a) expressed the view that researchers continue to use nil null hypotheses rather than non-nil null hypotheses for two reasons. First, most computer packages assume the researchers are testing nil null hypotheses. Thus, they are not equipped to invoke the necessary changes in calculations. As noted by Selin and Lapsley (1985, 1993), such changes include the use of critical values obtained from noncentralized t and F distributions. Second, some of the complexities of using non-nil null hypotheses are not yet readily applicable in many designs.

In spite of these two roadblocks, a testing technique is available to researchers who believe it is important to test non-nil hypotheses. Edgington (1995) has suggested that researchers can readily employ non-nil null hypotheses if they utilize randomization testing techniques. Edgington expressed the view that: “A randomization test null hypothesis need not
be simply one of no differential treatment effect [a nil null hypothesis] . . . but can . . . [reflect] response magnitudes [i.e., a non-nil null hypothesis]” (pp. 319-320).

**Acting on the Three Recommended Changes in Research Practices**

Our view regarding the debate on the need and importance of considering the statistical significance of the observed effect along with its effect size, i.e., practical significance, is more closely aligned with the position taken by Levin (1998), Levin and Robinson (2000), and Robinson and Levin (1997). Thus, we believe that statistical significance of the observed effect is an essential element to evaluate along with the practical significance of the effect.

The process we are encouraging educational researchers to use requires them to assess whether the observed effect is statistically significant. The process we are recommending, however, requires the observed effect not be statistically tested against the lack of any effect, i.e., a value of zero, but rather against a level deemed to be practically significant. Thus, we are advocating the use of non-nil null hypotheses that incorporate a value defined by the researchers as indicating practical significance. This practically significant value should be derived by the researchers in a manner based on the position articulated by Kirk (1996) who stated that it is important not to sanctify effect size numbers such as Cohen’s (1988) .2, .5, and .8. Kirk suggests that if practical significance is to be a useful concept, its determination must not be ritualized. Judgment regarding what is required for practical significance should inevitably involve a variety of considerations, including societal concerns and costs versus benefit comparisons, just to mention a few. We do believe, however, that Cohen’s identified effect size values, labeled as small, medium, and large, can be used by researchers and practitioners as a starting point for the identification of effect sizes that have meaning to them.
Since we are suggesting that non-nil null hypotheses should be employed, we recommend that it be statistically tested with randomizations tests. A randomization test has two desirable characteristics when used to test non-nil null hypotheses in educational and psychological settings. First, a randomization test will generate the distribution needed by the researcher to determine if the test statistic is statistically significant. Thus, the researcher would not need to incorporate special critical values as may be required by the use of the t and F values generated by most standardized statistical programs. Second, a random sample is not required when a randomization test is conducted. Edgington (1995) stated the position that: “A randomization test is valid for any kind of sample, regardless of how the sample is selected. This is an extremely important property because the use of nonrandom samples is common in experimentation” (p.6).

**Technique**

Our suggested analytic technique incorporates three major elements. First, a non-nil null hypothesis and its alternative are employed. Second, the key value contained in each of the hypotheses is one that is deemed as practically significant. Third, the non-nil null hypothesis is tested with a randomization test. As previously stated, the procedure we are recommending assumes that the concepts of statistical and practical significance are both essential components of the evaluation process. In addition, the value that is defined to be practically significant is thoughtfully determined by the researchers and practitioners through a process that may involve non-quantitative methods and becomes the key value in the non-nil null hypothesis.
An Illustration of the Recommended Testing Procedure

We believe the best way to present our suggested testing procedure is through its application to a research question and the corresponding data. In this illustration, we are using data collected in a study conducted by Piirto, Beach, Cassone, Rogers, and Fraas (2000). In this study, the authors were interested in determining whether high-school aged gifted students have higher intellectual scores than high-school aged non-gifted students. The intellectual scores measured the students’ levels of desire for knowledge and inquiry. This illustration includes a total of 49 gifted students and 51 non-gifted students. Each intellectual score was multiplied by 100 to facilitate the presentation of this illustration.

Before the non-nil null and the alternative hypotheses were constructed, the degree of difference between the mean scores of the gifted and non-gifted students, which was deemed by the researchers and practitioners to be practically significant, had to be established. After considerable discussion and reflection, a difference in the group means that exceeded four points was deemed necessary for practical significance to be achieved. The implications of the difficulty we encountered in arriving at this value will be presented in a later portion of this paper.

Since practical significance was equated with values in excess of four points, the non-nil null and alternative hypotheses were constructed as follows:

$H_0$: The mean of the gifted students does not exceed the mean of the non-gifted students by more than four points.

$H_1$: The mean of the gifted students does exceed the mean of the non-gifted students by more than four points.
This non-nil null hypothesis was tested with a randomization test, which was generated by a computer program entitled Resampling Stats Add-In for Excel (Blank, Seiter, & Bruce, 1999). As discussed in a previous work (Fraas & Newman, 2001), we believe most researchers will find this macro add-in software easier to use than the other software produced by the same company, which is entitled Resampling Stats (Simon, Weidenfeld, Bruce, & Puig, 1999).

Before the students' scores were subjected to the randomization test, the value of four was subtracted from each gifted student's score. The mean of the gifted students in the sample was 23.65. Thus, the mean of the gifted students' modified scores was 19.65. The standard deviation of the mean of the gifted students was 16.04 which, of course, matched the standard deviation of the modified scores. The mean and standard deviation values for the non-gifted students were 15.55 and 14.75, respectively.

Once the scores of the gifted students and the non-gifted students were entered into the randomization test program, it generated a distribution of 10,000 differences between the modified mean of the gifted students and the mean of the non-gifted students. The program calculated the proportion of the 10,000 values in the distribution that exceeded the difference between the modified mean of the gifted students ($\overline{x} = 19.65$) and the mean of the non-gifted students ($\overline{x} = 15.55$). This proportion (.094), was compared to an established maximum proportion of values in the distribution that we were willing to obtain and still reject the non-nil null hypothesis, which was our alpha level. We subjectively established this maximum proportion to be .05. Since the calculated proportion of .094 exceed the established maximum proportion of .05, we were not willing to reject the non-nil null hypothesis. Thus, we concluded that any difference between the means of the gifted students and the non-gifted students in
excess of four points, was more likely to occur by chance at a level greater than we were willing
to accept.

**Difficulty in Establishing the Value for Practical Significance**

Implementing this procedure revealed to us that establishing the level beyond which
researchers and practitioners would consider the difference to be practically significant is not a
simple task or one with which researchers and practitioners have had a great deal of experience.
In this application we found ourselves relying on Cohen’s low, medium, and high guidelines.
We believe this task should be qualitative in nature and the reliance on Cohen’s guidelines is not
the best practice to follow. In regards to this point we are in agreement with Kirk (1996) who
stated:

> With respect to determining the practical significance of results, Cohen’s definitions of
small, medium, and large effects represent a good beginning. However, much more
systematic research is needed to extend his work. . . . If practical significance is to be a
useful concept, its determination must not be ritualized. (p. 756)

The development of procedures and thought processes that could assist researchers and
practitioners with the task of identifying practical significance levels may prove very beneficial to
the field of research methodology.

**Summary**

We have proposed a hypothesis testing technique that utilizes a non-nil null hypothesis in
which the key value is set by the researchers and practitioners at a level deemed necessary for
practical significance to be reached. Once the non-nil null hypothesis is constructed, it is tested
by means of a randomization test. This technique stresses two key philosophical positions. First,
the concepts of practical and statistical significance are both essential components in the evaluation process. Second, the use of this procedure allows the researchers to consider a variety of criteria beyond statistical ones when establishing practical significance levels.

During the process of implementing this testing procedure we discovered one task in which further study, reflection, and possibly, training of researchers may prove fruitful. That task is the establishment of the practical significance level. When we applied this technique to the overexcitability scores, the practitioners and researchers found the establishment of a practical significance level to be a rather nebulous one. We found ourselves relying to a significant extent on Cohen’s effect size guidelines. We believe that further developments in the methods used to identify practical levels of significance and training in those methods for practitioners and researcher alike may prove to be very important for future research and evaluation projects.
References


Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology* 9 (2), 165-181.