3 *t*-Test: The Good, the Bad, the Ugly, & the Remedy

Guili Zhang
East Carolina University
Greenville, North Carolina

ABSTRACT

The *t*-test is one of the most commonly used significance tests to assess whether the means of two groups are statistically significantly different from each other. The use of *t*-test has become a natural choice, and rarely practitioners question its appropriateness. This article reviews and discusses *t*-test's value in providing a rough comparison of two means (the good), its inability to provide the amount or magnitude of the difference in means (the bad), and its deficiency in that its outcome essentially depends completely on the sample size (the ugly). The *t*-test's failure to provide the amount of difference offers no basis for us to judge practical significance. Furthermore, it makes meta-analysis impossible or ineffective and, therefore, seriously hinders the advancement of educational research. Its direct dependency on sample size makes a *t*-test virtually useless in providing reliable information. Effect size (ES) and confidence interval (CI) for ES are recommended to address *t*-test's inadequacies by serving as a supplement to or replacement of the *t*-test. The commonly used effect size indices are reviewed and discussed in regard to their performance and robustness.

INTRODUCTION

Middle-grades researchers often use *t*-tests to determine if two means are statistically significantly different from each other. Researchers are happy when the *t*-test informs them that group A's mean is bigger than group B's or vice versa. Researchers rarely ask themselves: Is this all I can know? Do I not want to know how much bigger group A's mean is than group B's? Furthermore, they often fail to realize that the statistical significance of the difference between the two groups' means directly depends on the sample size. In fact, given big enough sample sizes the two means will almost always be statistically significantly different from each other!

It has been nearly 15 years since the publication of Cohen's wakeup call article (Cohen, 1994) titled *The Earth is Round ($p < .05$),* in which he discussed 4 decades of severe criticism of significance testing and rigorously criticized this practice including the use of the *t*-test. Yet, significance testing, the mechanical dichotomous decision around a sacred .05 criterion, still persists, despite the great deal of mischief that has been associated with the test of significance (Bakan, 1966).

Using an ES in addition to or in place of a hypothesis test such as a *t*-test has been recommended by some statistical methodologists since as early as the 1960s because ESs are recognized as being more appropriate and more informative (Cohen, 1965; 1994; Cumming & Finch, 2005; Finch, Thomason, & Cumming, 2002; Hays, 1963; Meehl, 1967; Nickerson, 2000; Steiger & Fouladi, 1997). Further, a CI for ES contains all the information found in the significance test as well as vital information not provided by the significance test about the magnitude of effects and precision of estimates (Cohen, 1994; Cumming & Finch, 2001; 2005). A CI indicates the range of population ESs with which the data are consistent. By contrast, a hypothesis test merely indicates whether the data are consistent with a population ES of zero or not.

PERSPECTIVE & METHODOLOGY

t-test the Good: It Tells the Direction

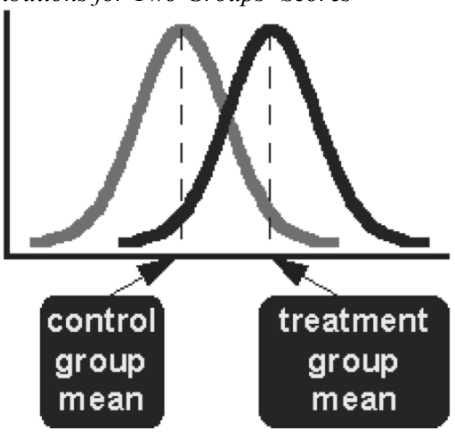
The *t*-test is often used to test a hypothesis regarding the equality of two means. The process is as follows (Fraenkel & Wallen, 2008):

1. Specify the research hypothesis, which is the predicted outcome of a study (e.g., There is a difference between the population mean of students using method A and that of students using method B).
2. State the null hypothesis (H_0) (e.g., There is no difference between the population means of students using method A and students using method B).

3. Find the sample statistics pertinent to the hypothesis.
4. Find the *p*-value, that is, the probability of obtaining the observed results in the samples if the null hypothesis is true. This is done by comparing the observed *t* statistic with the *t* distribution. The shape of the *t* distribution is directly tied to the sample size, or more preciously the degree of freedom, *df*, which is the number of *independent* observations.
5. Reject H_0 and accept the research hypothesis if the *p*-value is smaller than the pre-specified type I error rate α, usually .05.
6. Do not reject H_0 if the *p*-value is greater or equal to α.

As such, the *t*-test serves as an administratively convenient method through which researchers obtain crude information about the comparison of two means. Specifically, with a non-directional hypothesis, the *t*-test answers the question, "are the two population means different?", and in a directional hypothesis, it answers the question regarding the direction of the difference, "is the mean of population A bigger or smaller than that of population B?" Such information, although rudimentary, can be useful in some situations.

Figure 1
Idealized Distributions for Two Groups' Scores



t-test the Bad: It Does Not Tell Us the Magnitude of Difference

Significance testing has not only failed to support the advance of psychology (social science) as a science, but also has seriously impeded it (Cohen, 1994). What it tells researchers is very elementary and crude. It tells the direction (i.e., which one is bigger?), not the amount, that is, how much group A's mean is bigger than group B's. Tukey (1991) discounted the value of what a *t*-test can offer researchers by stating "It is foolish to ask 'are the effects of A and B different?' They are always different—for some decimal place" (p. 100).

The ritual dichotomous reject-accept decision is not the way any science is done (Cohen, 1994). Cohen's opinion echoed Rozeboom's assertion in as early as 1960, "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one . . . believes the hypothesis. . . being tested" (p. 420). Physical scientists have learned much by storing up the amounts, not just directions. Measuring things on a communicable scale lets researchers stockpile information about amounts, which serves as foundation for meta-analysis.

t-test *the Ugly: It Completely Depends on Sample Size*

The ugly side of the *t*-test is, given big enough sample size, the null hypothesis is always false (Cohen, 1994). This shortcoming very much discredits the value of *t*-test entirely. As Cohen (1990) stated,

> It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So, if the null hypothesis is always false, what's the big deal about rejecting it? (p.1308)

In fact, Berkson expressed a similar view in as early as 1938, "It would be agreed by statistician that a large sample is always better than a small sample…But since the result of the former [large sample] test is known, it is no test at all" (p. 526f). Thompson (1992) made the same point even more piercingly:

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. This tautology has created considerable damage as regards the accumulation of knowledge (p. 436).

t-test: The Remedy

An important caveat when overusing *t*-tests (see Capraro & Capraro in this volume) as a first option for avoiding multiple univariate tests is to use a multivariate method like canonical correlation analysis (see Yetkiner elsewhere in this volume). However, when considering *t*-tests, effect sizes are recognized as being more appropriate and more informative than *t*-tests (Cohen, 1965, 1994; Cumming & Finch, 2005; Finch et al., 2002; Hays, 1963; Meehl, 1967; Nickerson, 2000; Steiger & Fouladi, 1997). In the last two decades,

reporting an ES has become mandatory in some editorial policies (Murphy, 1997; Thompson, 1994) and is strongly recommended for *American Psychological Association (APA)* journals. *The Publication Manual of the American Psychological Association* (2001) stated that it is almost always necessary to include some index of ES or strength of relationship in the results section of a research paper. In fact, the manual cites the failure to report effect sizes as a defect in reporting research. The APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999) also supported this position. A CI for an ES is recommended as a superior replacement for significance testing because this CI contains all the information found in the significance tests and vital information not provided by the significance tests about the magnitude of effects and precision of estimates (Cohen, 1994; Cumming & Finch, 2001, 2005).

The merit of providing magnitude of effects and precision of estimation is clearly shown in the following example. Suppose a researcher conducted two experiments to test whether there was a significant difference between the treatment group and the control group. Recall that Cohen (1969) suggested .20, .50, and .80 as small, medium, and large ESs in experiments comparing two treatments. There were 400 participants in experiment A, resulting in 200 participants in each group. The researcher found a small ES of .20 which was significantly different than zero ($t = 2$, $p < .05$) and indicated statistical support for a treatment effect. For experiment B, the researcher was only able to find 16 participants, resulting in 8 participants in each group, and found a large ES of .80, which was not significantly different from zero ($t = 1.6$, $p > .05$) and did not indicate support for a treatment effect. Given hypothesis testing and effect sizes, one apparent contradiction is that a small ES results in statistically significant evidence of a treatment effect (experiment A), while a large ES results in non-significant evidence of a treatment effect (experiment B).

The apparent contradiction can be resolved by setting a CI for the ES in each experiment. In experiment A, the CI is [.003, .396]. The CI tells us that the data in experiment A are consistent with a population ES between .003 and .396, and thus indicates a population ES that is larger than zero but may be extremely small or may approach a medium ES. In experiment B, the CI is [−.236, 1.810], indicating the data are consistent with a population ES anywhere between –.236 and 1.810 and thus indicates that the population ES may be anywhere from a small ES in favor of the control treatment to a very large ES in favor of the experimental treatment. Clearly the CIs provide more complete information than the ES and the hypothesis test. The CIs also explain why the hypothesis tests indicated that a small ES was significantly different from zero in experiment A, while a large ES was not significantly different from zero in experiment B. In Experiment A, the

population ES was much more precisely estimated than the population ES in experiment B.

EFFECT SIZE INDICES

A large number of ES indices have been developed and proposed (Algina, Keselman, & Penfield, 2005). There are a variety of commonly used ESs measuring separation of two independent samples among them are Cohen's *d* (Cohen, 1965) (for more information about Cohen's *d, see the article by Thompson in this issue)*, Glass's *d*, Hedges' *g* (Hedges & Olkin, 1985), two versions of Cohen's *d* based on trimmed means and Winsorized variances ($d_R^{\dagger}$ suggested by Hogarty & Kromrey, 2001 and d_R suggested by Algina et al., 2005), eta squared, omega squared, McGraw and Wong's (1992) common language ES (CL), Cliff's dominance statistic (1993, 1996), Kraemer and Andrews γ_1^* (1982), Wilcox and Muska's *W* (1999), and Vargha and Delaney's *A* (2000).

Hedges' *g* is a variation of Cohen's *d* that provides some adjustment to reduce small sample bias. Eta squared and omega squared are similar in that both are proportion-of-variance measures conveying the proportion of the total variance that is due to the groups. McGraw and Wong's (1992) CL provides an estimate of the probability that a score for a member in a population is greater than a score for a member in another population under the assumption of independence, normality, and equal variances. Denoting this probability as $p_{2>1}$, Cliff's dominance statistic estimates the value of $p_{2>1} - p_{1>2}$. In other words, Cliff's *d* is used to estimate the degree to which the probability that a score in the second treatment exceeds a score in the first treatment is different from the probability that a score in the first treatment exceeds a score in the second treatment, the property that Cliff referred to as "dominance." Kraemer and Andrews' γ_1^* (1982) is a nonparametric index of effect size that is based on the degree of overlap between samples. Wilcox and Muska's *W* is a nonparametric analogue of omega squared that estimates the degree of certainty with which an observation can be associated with one population rather than the other. Vargha and Delaney's *A* is a measure of stochastic superiority that applies to distributions that are at least ordinally scaled (Hogarty & Kromrey, 2001).

Performance of the Effect Size Indices

Research investigating the performance of the various ES measures is fairly limited. Hedges and Olkin (1985) suggested that Cohen's *d* evidences a small sample bias. Hogarty and Kromrey (2001) compared the performance of nine effect size indices when they are used in the context of populations with various levels of nonnormality and variance heterogeneity. The nine indices include Cohen's *d*, Cliff's dominance statistic, $g, \gamma_1^*, CL, A, d_R^\dagger$, a naïve estimator of *W* and a .632 bootstrap estimator of *W*. They evaluated the performance of these nine indices in terms of their statistical bias and standard error by using Monte Carlo methods. The results indicate that Cohen's *d* and Hedges' *g*, the two most frequently used effect size estimates, showed nontrivial sensitivity to violations of normality and homogeneity of variance, which confirms the concerns raised about the appropriateness of using these indices as indicators of effects in such populations (Kraemer & Andrews, 1982; Wilcox & Muska, 1999). In addition, $d_R^\dagger$ evidenced severe bias under small sample conditions. Indices *CL*, γ_1^* and the naïve estimator of *W* only appeared to be slightly less sensitive than Cohen's *d* and Hedges' *g,* but showed pronounced bias under small sample size condition or nontrivial sensitivity to violations of normality and homogeneity of variance. Cliff's dominance statistic and Vargha and Delaney's *A* showed better performances in producing relatively unbiased estimates and consistent standard errors.

Hess and Kromrey (2004) investigated the performance of the CIs for Cohen's *d* and Cliff's dominance statistic constructed by using seven CI construction methods. The CI construction methods included the normal theory *Z* band, the percentile bootstrap, the bias corrected bootstrap, the bias corrected and accelerated bootstrap (BCa), pivotal, Studentized pivotal, and the Steiger and Fouladi interval inversion band. (For a description of these methods, see Hess & Kromrey, 2004). Monte Carlo methods were used to compare confidence interval estimates using random samples generated from populations under known and controlled conditions. There were four design factors: (a) sample size ranging from 5 to 200 including balanced and unbalanced designs, (b) population effect sizes .00, .20, .50, and .80, (c) population distribution shape (population skewness and kurtosis of 0,0 and 2,6, and (d) variance in the two populations: 1:1, 1:2, 1:4, and 1:8. Across all of the conditions, all of the CI construction methods provided slightly better coverage probabilities for Cliff's dominance statistic than for Cohen's *d*, with the exception of the Pivotal Bootstrap method.

Reporting a CI for the effect size is important as was well put by Wilkinson et al. (1999) (for more information on CI's consider the articles by Capraro & Capraro and Thompson in this issue). Steiger and Fouladi (1997) asserted that "it is much more informative to state a confidence interval on $\mu_2 - \mu_1$ than it is to give the *p* level for the *t*-test of the hypothesis that $\mu_2 - \mu_1 = 0$. Interests in the accuracy and usefulness of the ESs have motivated explorations of the usefulness and effectiveness of CIs for ESs (cf. Algina & Keselman, 2003a, 2003b; Cumming & Fitch, 2001). For more information on CI's consider Skidmore elsewhere in this volume.

RESULTS & DISCUSSION

It is recognized that *t*-test is not an effective method for comparing two means. The *t*-test only informs us whether "A is different from B." Moreover, this only piece of information that *t*-test provides us is virtually worthless because given big enough sample size, A is always different from B. Therefore, we cannot learn much from a *t*-test. It is clear that amount, as well as direction is vital.

Confidence intervals for effect size are strongly recommended to be used as a useful supplement to and maybe even a superior replacement for the *t*-test. Effect size indices such as Cohen's δ are able to provide all information that is provided by the *t*-test as well as vital information not provided by the *t*-test such as the magnitude of the effects and precision of estimates.

The use of adequate effect size index and confidence intervals to supplement or in place of *t*-test will make studies on two group comparisons more informative and useful as a basis for judging practical significance. Quantifying the differences between the means on a commutable scale also makes meta-analysis possible and the replication and comparisons of research studies achievable, which can serve as the cornerstone for the advancement of quantitative educational research.

REFERENCES

Algina, J., & Keselman, H. J. (2003a). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, *63*, 537-553.

Algina, J., & Keselman, H. J. (2003b, May). *Confidence intervals for effect sizes*. Paper presented at a conference in honor of H. Swaminathan, Amherst, MA.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). Effect sizes and their intervals: The two-level repeated measures case. *Educational and Psychological Measurement, 65*, 241-258.

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

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 1-29.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics, *Multivariate Behavioral Research, 31*, 331-350.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.

Cohen, J. (1969). *Statistical power analyses for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1302.

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

Cumming, G., & Finch. S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 178-180.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 36*, 312-324.

Fraenkel, J. R., & Wallen, N. E. (2008). *How to design and evaluate research in education*. San Francisco: McGraw-Hill.

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart & Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hess, M. R., & Kromrey, J. D. (2004, April). *Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under non-normality and heterogeneous variances.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletion, 91*, 404-412.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletion, 111*, 361-365.

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

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3-5.

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

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

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? Hillsdale, NJ: Erlbaum.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*, 101-132.

Wilcox, R. R., & Muska, J. (1999). Measuring effect size: A non-parametric analogue of ω^2. *British Journal of Mathematical and Statistical Psychology, 52*, 93-110.

Wilkinson, L., & the Task force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.