Statistical Significance and Effect Size in Education Research: Two Sides of a Coin

XITAO FAN
University of Virginia

ABSTRACT **In education research, statistical significance and effect size are 2 sides of 1 coin; they complement each other but they do not substitute for each other. Good research practice requires that, to make sound research decisions, both sides should be considered. In a simulation study, the sampling variability of 2 popular effect-size measures (d and R^2) was examined. The variability showed that what is statistically significant may not be practically meaningful, and what appears to be practically meaningful could have been the result of sampling error, thus not trustworthy. Some practical guidelines are suggested for combining the 2 sources of information in research practice.**

Key words: effect-size measures, research practice, statistical significance testing

In education research, statistical significance testing has received many valid criticisms in recent years primarily because the outcome of statistical significance testing relies too heavily on sample size, and the issue of practical significance is often ignored. Consequently, too much reliance on statistical significance testing often limits understanding and applicability of research findings in education practice. Effect size has been proposed as a supplement or an alternative to statistical significance testing; it has become increasingly popular. Some education researchers, however, may not be aware that, by itself, effect size can also be misleading because sample size influences the sampling variability of an effect-size measure. Through a Monte Carlo experiment, I show that statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in education research, both sides should be considered. To lay a foundation for the discussion in this article, I first reviewed some major issues related to statistical significance testing and effect-size measures.

Statistical Significance Testing

Use of statistical significance testing in research. There have been different misconceptions about what significance testing is and what it is not (Shaver, 1993). For this article, one should have a good understanding about the basic purpose of statistical significance testing in quantitative research and about what information statistical significance testing provides for education researchers.

The fundamental concept underlying statistical significance testing is sampling variation. From a population with a known parameter (e.g., known population mean), sample statistics (e.g., observed means of multiple samples) will vary around the population parameter to a certain extent. Because of the sampling variability, the difference between an observed sample statistic (e.g., sample mean) and the population parameter (i.e., population mean) does not necessarily indicate that the sample does not belong to the population. For example, if the mean of a random sample (N = 20) is observed to be 68, could this sample statistic have occurred because of sampling variability (i.e., by chance) if the population mean is 80? A statistical significance test can be conducted to evaluate the viability of the hypothesis that the sample with a mean of 68 could have been drawn from a population with a mean of 80. That evaluation is done by assessing how likely the difference between the observed sample statistic and the known population parameter could have occurred as the result of chance, that is, random sampling variation. In other words, statistical significance testing answers the question: What is the probability of obtaining an observed sample statistic (e.g., mean of 68) when the population has a known parameter value (e.g., population mean of 80)?

Address correspondence to Xitao Fan, University of Virginia–Curry School of Education, 287 Ruffner, 405 Emmet Street South, PO Box 400277, Charlottesville, VA 22904-4277. (E-mail: xfan@virginia.edu)

Assume that two treatment conditions (A and B) exist; (e.g., A represents a new instructional approach for teaching mathematics and B represents the conventional instructional approach currently in use). The researcher is interested in knowing if A is more effective than B in teaching mathematics (RH: research hypothesis, A is better than B). To help decide if the RH can be supported, the researcher set up another hypothesis (NH: null hypothesis of no difference) that A and B are equally effective, that is, students under A and B will learn equally well.

Because of sampling variation, even if A is the same as B in terms of effectiveness, the sample under A may have higher sample mean than the sample under B. So the observation that students under A performed better than those under B does not necessarily mean that Method A is more effective than B because sampling error has not been ruled out as one possible explanation for the observed difference between the two samples.

If A and B treatments are the same (NH: of no difference), a small performance difference between A and B samples is more likely to occur by chance than is a large performance difference. When the difference between the two samples becomes sufficiently large relative to the random sampling variation, however, one begins to doubt that A and B are equally effective. In that case, it should be highly unlikely to observe the large performance difference between the two samples. The question becomes: How much higher should the mean of the Method A sample be than that of the Method B sample before one can determine with reasonable confidence that the observed difference is not due to sampling variability (i.e., chance)? Once one decides statistically that the sampling variability is no longer a viable explanation for the observed difference, the NH will be rejected in favor of the RH (Method A is more effective than Method B). The rejection of NH constitutes evidence for supporting the RH because a statistical significance test helps to eliminate sampling error, or chance, as a viable explanation for the observed difference between the two samples.

In the statistical significance testing, I assessed the probability of obtaining the sample data (D) if the null hypothesis (H_0) is true, that is, $p(D \mid H_0)$. If $p(D \mid H_0)$ is sufficiently small (e.g., smaller than .05 or .01), the null hypothesis will be considered not viable and will be rejected. The rejection of the null hypothesis indicates that the random sampling variability is the unlikely explanation for the observed statistical results, but it does not generally show the importance of obtained statistical results. Regarding the example of A and B methods in teaching mathematics, rejection of the null hypothesis (A and B are equally effective in teaching mathematics) simply means that, given the observed magnitude of difference between the two samples, it is highly unlikely that sampling error could have been the cause for the observed difference. As a result, one concludes that A and B are probably not equally effective. That conclusion, however, does not provide a clear indication about how much more effective Method A is than Method B in the practical sense. Unfortunately, the meaning of statistical significance testing as discussed here has sometimes been lost, and the importance of statistical significance tends to be grossly exaggerated in education research practice.

Major criticisms of statistical significance testing. In research and evaluation studies, the overreliance on statistical significance testing has been challenged on several grounds. Thompson (1993) discussed three relevant criticisms for statistical significance testing: (a) overdependency on sample size, (b) some nonsensical comparisons, and (c) some inescapable dilemmas created by statistical significance testing (e.g., testing for assumption vs. testing for the research hypothesis). In a similar vein, Kirk (1996) discussed three major criticisms of statistical significance testing: (a) Significance testing does not tell researchers what they want to know, but rather, it creates the illusion of probabilistic proof by contradiction (Falk & Greenbaum, 1995). (b) Statistical significance testing is often a trivial exercise because it simply indicates the power of the design (which primarily depends on the sample size) to reject the false null hypothesis. (c) Significance testing "turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision," and this dichotomous decision process may "lead to the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions" (Kirk, p. 748) because of the slight differences in their design (e.g., sample sizes).

Of all the criticisms for statistical significance testing, the best known is probably its overreliance on sample size. It is well known that the outcome of statistical significance testing depends heavily on the sample size used for the testing: For a fixed amount of difference between the hypothesized population parameter and the observed sample statistic, the larger the sample size, the easier it is to reject the null hypothesis. As discussed by Meehl (1978), ". . . the null hypothesis, taken literally, is always false" (p. 822), and because of this, statistical significance often becomes a matter of having a sufficiently large sample in order to have enough statistical power for rejecting the null hypothesis. Thompson (1992) sarcastically commented that, in the ritualistic exercise of significance testing, ". . . tired researchers, having collected data from 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're tired" (p. 436).

Because the importance of statistical significance testing traditionally has been exaggerated, it has become something sacredly ritualistic in quantitative research, to the point that statistical significance almost becomes the literal equivalent of the importance of quantitative findings. Undoubtedly, that misconception has been compounded by the unfortunate misnomer of "significance" in this context.

Effect Size

The criticisms of statistical significance testing have led education researchers to explore other approaches for mak-

ing quantitative sense out of the data because, as reasoned by many researchers (e.g., Kirk, 1996), the rejection of the null hypothesis by itself is not very informative in the practical sense. There is little doubt that the importance attributed to statistical significance testing in education research practice has traditionally far exceeded what it really deserves (e.g., Thompson, 1993).

Use of effect-size measure. Because statistical significance shows only in probabilistic terms how unlikely it is to obtain the sample data if the null hypothesis is true but does not inform whether the findings are practically meaningful or important, the general approach of obtaining some kind of scale-free effect-size measure as the indicator of practical meaningfulness or importance has become popular, and its use in research practice has been advocated widely in recent years (Thompson, 1996). As the *Publication Manual of the American Psychological Association* (APA) (4th edition) explains, neither a priori nor exact probabilistic values reflect "the importance (magnitude) of an effect or the strength of a relationship because both probability values depend on sample size. You can estimate the magnitude of an effect with a number of measures that do not depend on sample size" (American Psychological Association, 1994, p. 18). Recently, The APA Task Force on Statistical Inference admonished, "Always provide some effect-size estimate when reporting a p value" (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 599).

Although there is some consensus that the role of statistical significance testing in research should be reduced, there is less agreement about the extent to which it should be reduced and about the extent to which the role of effect size should be enhanced in quantitative research. On one hand, statistical significance testing has been criticized as representing almost nothing but obstacles for scientific inquiry (Carver, 1978; Meehl, 1978), as indicated by the strongly worded criticism that the reliance on significance testing for the null hypothesis "is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl, p. 817).

On the other hand, some researchers have defended the legitimate role that the correct use of significance testing plays in scientific inquiry (Levin, 1993; Schafer, 1993). Levin argued that the *baby* (statistical significance testing) should not be thrown out with the *bath water,* just because the bath water might not be clean (misuse/misinterpretation of significance testing). Levin used hypothetical examples to argue that, even with effect-size measures, statistical significance testing is still essential in many situations so that researchers are not misled by effect-size measures.

Effect-size measures. Different measures for effect size have been developed over the decades. Both Kirk (1996) and Snyder and Lawson (1993) provided useful and practical summaries of those measures. Because the terminology used for describing the variety of effect-size measures has not been standardized in the literature, confusion sometimes occurs about what effect-size measure has been reported in a study (Kirk, 1996). Maxwell and Delaney (1990) categorized the variety of effect-size measures into two broad categories: measures of effect size (according to group mean differences) and measures of association strength (according to proportion of variance accounted for).

The first category, measures of effect size, is based on standardized group-mean difference. That category of measures is represented by Cohen's d or some variations of it (e.g., Glass's g for meta-analysis and Hedges' g). In the most general form, d is expressed as follows:

$$d = \frac{\bar{X}_{\text{group 1}} - \bar{X}_{\text{group 2}}}{SD_{\text{pooled}}}$$

where SD_{pooled} represents the pooled standard deviation between the two groups. In research situations in which two groups are involved and the comparison of the group means is the primary interest, d has become the measure of choice for effect size.

The second broad category, measures of association strength, is based on the proportion of variance accounted for, and it can be represented by R^2 or η^2. The most general form for the association strength can be expressed as follows:

$$\eta^2 = \frac{\text{Sum of squares}_{(\text{a source})}}{\text{Sum of squares}_{(\text{total})}}$$

The numerator represents the sum of squares from a source of interest. In a model that contains one explanatory factor (predictor) only, the source of interest is obviously the only explanatory variable in the model. In that case, R^2 is the usual term used as the measure of association strength. For a model containing multiple factors (predictors), the source of interest may include either a subset of factors (e.g., sum of squares due to one predictor from a multiple-predictor model) or all the explanatory factors (predictors) in the model. In the former case, η^2 is typically the term used to describe the proportion of variance accounted for by a subset of factors (predictors). In the latter case, R^2 is usually used as the term to describe the proportion of variance accounted for by all the factors (predictors) in the model (i.e., the full model). In that sense, η^2 and R^2 are conceptually the same.

Because R^2 contains upward bias due to the maximization property of the least-square principle, different bias-corrected counterparts of R^2 have been proposed, such as ω^2, ε^2, and others (see computational details in Kirk, 1996, and Snyder & Lawson, 1993). A literature review of several influential journals in psychology has shown that R^2 is the most popular measure reported for measuring association strength (probably because of its availability from statistical software programs), whereas bias-corrected counterparts of R^2 (e.g., ω^2, ε^2, and others) have been minimally reported (Kirk, 1996).

Effect size as a random variable. Many researchers seem to have ignored an important aspect of effect-size measure

when they used a sample effect size in research: Effect-size measure is a random variable just as sample mean is a random variable. Being a random variable has one important implication for its interpretation: Sample effect-size measure is subject to sampling variability as dictated by its underlying sampling distribution. Furthermore, the extent of sampling variability of an effect-size measure is affected by sample size, similar to the situation in which the probability associated with statistical significance testing is influenced by the sample size. In other words, when the sample size is small, the sample effect size may deviate farther from the population effect size than when the sample size is large.

Although the random variable nature of effect-size measures has been widely known in the quantitative literature (e.g., Fowler, 1985; Glass & Hopkins, 1996, chapter 14; Hedges & Olkin, 1985), relatively few education research practitioners pay sufficient attention to, or show enough interest in, this knowledge. In education research literature, it is not uncommon to encounter discussion of the effect that the outcome of a statistical significance test is influenced by sample size (true), so attention should focus on effect size as if it were not influenced by sample size. Undoubtedly, the use of effect-size measure makes good quantitative and common sense; but education researchers should realize that the use of effect size serves a different purpose than that of a statistical significance test. Whereas statistical testing evaluates the probability of obtaining the sample outcome by chance (sampling error), effect size provides some indication for practical meaningfulness. Although a statistically significant outcome may not be practically meaningful, a practically meaningful outcome may also have occurred by chance, and, consequently, is not trustworthy.

The general purpose of this article was to demonstrate that both statistical significance testing and effect size are needed to make sound research decisions. Because the two items serve different purposes, they supplement each other, but do not substitute for one another. To accomplish the general goal of this article, I addressed the following specific objectives:

1. To empirically assess the extent of sampling variability of major effect-size measures
2. To empirically assess the effect of sample size on the variability of sample effect sizes
3. To offer some practical guidelines in combining the statistical significance test outcome and the descriptive effect-size measure to reach sound quantitative decisions in research

Method

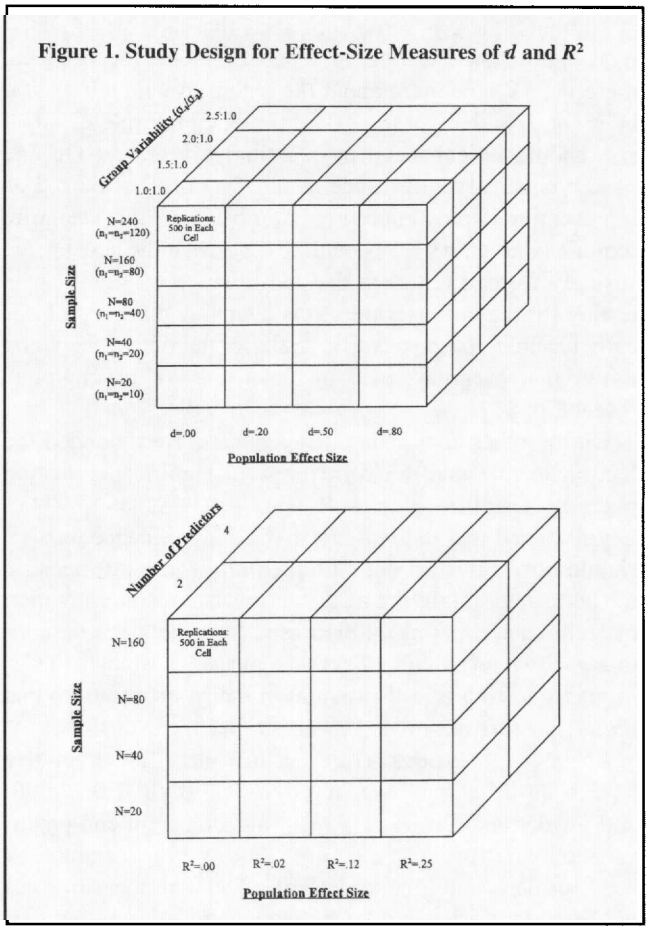
Although theoretical sampling distributions of some popular effect-size measures have been known (e.g., see Hedges & Olkin, 1985 for d, Glass & Hopkins, 1996 for R^2), I used an empirical approach in this article to provide more intuitive discussion about the relevant issues. I conducted a Monte Carlo experiment to simulate different data conditions under which both effect-size measures and statistical significance testing outcomes were obtained and later analyzed.

Design

In this article, I used the two most widely known effect-size measures: d (standardized mean difference) and R^2 (proportion of variance accounted for). The two effect-size measures are generally known to researchers who have been exposed to the concept of effect size. The literature review of several psychology journals by Kirk (1996) indicates that R^2 is by far the most frequently reported effect-size measure, probably because it is routinely reported in regression or general linear-model procedures. The meta-analysis work by Glass (1976) undoubtedly contributed to the popularity of d as the effect-size measure.

I generated samples from two statistical populations with known population parameters to evaluate standardized two-group mean difference (d); see Figure 1. I considered three factors in the Monte Carlo simulation design: (a) four levels of population effect size (d = .00, .20, .50, and .80, respectively) that correspond to zero, small, medium, and large effects as suggested by Cohen (1988, chapter 2); (b) five levels of sample-size conditions (N = 20, 40, 80, 160, 240); and (c) four

Figure 1. Study Design for Effect-Size Measures of d and R^2



conditions of group variability ratio (represented by the population standard deviations) between two populations (σ_1/σ_2 = 1, 1.5, 2, and 2.5, respectively). For the fully crossed design, the three factors yielded 80 ($4 \times 5 \times 4$) cells. I conducted 500 replications within each cell; the total number of replications in this Monte Carlo experiment for evaluating d were 40,000 (500×80).

I used regression models to evaluate R^2 (proportion of variance accounted for). I considered three factors in the design: (a) four levels of population effect size (R^2 = .00, .02, .12, and .25, respectively), which approximately correspond to zero, small, medium, and large effects as suggested by Cohen (1988, chapter 9); (b) four levels of sample-size conditions (N = 20, 40, 80, and 160, respectively); and (c) two conditions for the number of predictors (k = 2 and 4, respectively), with the correlation among the predictors set at r = .10. The fully crossed design of the three factors called for 32 cells ($4 \times 4 \times 2$). With 500 replications within each cell, the total number of replications for the experiment was 16,000 (32×500). The designs for evaluating d and R^2 are presented graphically in Figure 1.

Data

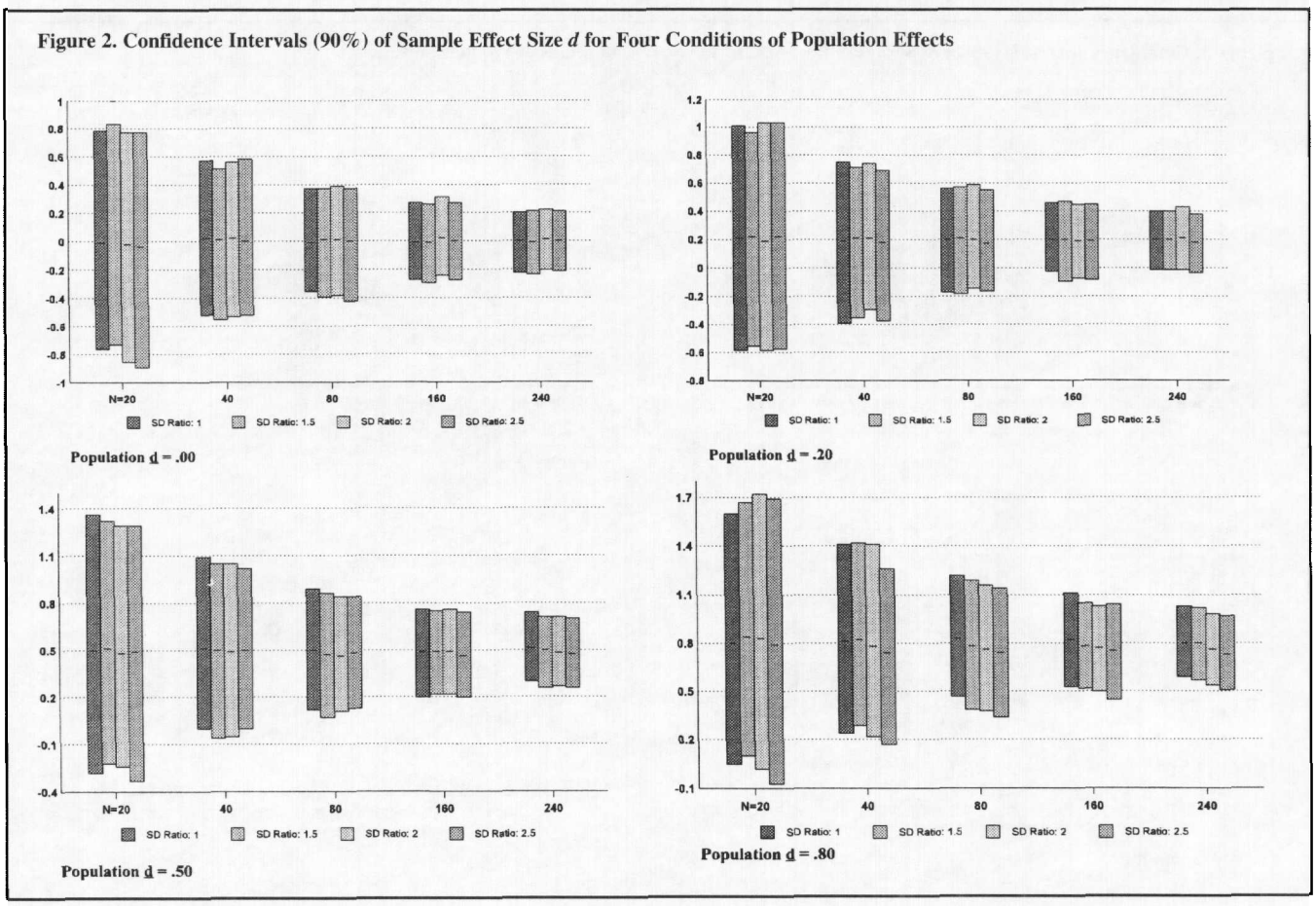
I attained data generation by using the SAS normal data generator. Multivariate normal data for regression models were simulated with the matrix decomposition procedure (Kaiser & Dickman, 1962). I accomplished all sample data generation, sample effect-size calculation, and statistical significance testing through the Interactive Matrix Language (PROC IML) of the SAS system (SAS Window Version 7.0). I did not consider data nonnormality in this study. As a result, the influence of data nonnormality on both effect-size measures and statistical significance test outcomes was not assessed.

Results and Discussion

Figure 2 graphically describes the sampling variability of the effect-size measure of d for four conditions of population effects: zero, small, medium, and large (population d = .00, .20, .50, and .80, respectively). In addition to sample-size conditions, the four conditions of group variability ratio(σ_1/σ_2) are also presented (σ_1/σ_2 =1, 1.5, 2, and 2.5). In Figure 2, a high–low bar represents the 90% confidence interval of sample d for a condition of sample size and for a group variability ratio (ratio of the standard deviations of two groups), and a short horizontal line within a bar represents the mean of 500 sample ds.

Several observations can be made from Figure 2. First, sample effect-size measure d appears to be an unbiased[1] estimate of population d. The lack of bias of sample d is

Figure 2. Confidence Intervals (90%) of Sample Effect Size *d* for Four Conditions of Population Effects



obvious because over repeated sampling, the mean of sample d is very close to the known population value specified in the Monte Carlo experiment (population d = .00, .20, .50, and .80, respectively) under most data conditions. However, a larger discrepancy between the two population standard deviations (*SD* ratio) causes some minor degree of downward bias of sample d, and this is especially obvious under the condition of population d = .80.

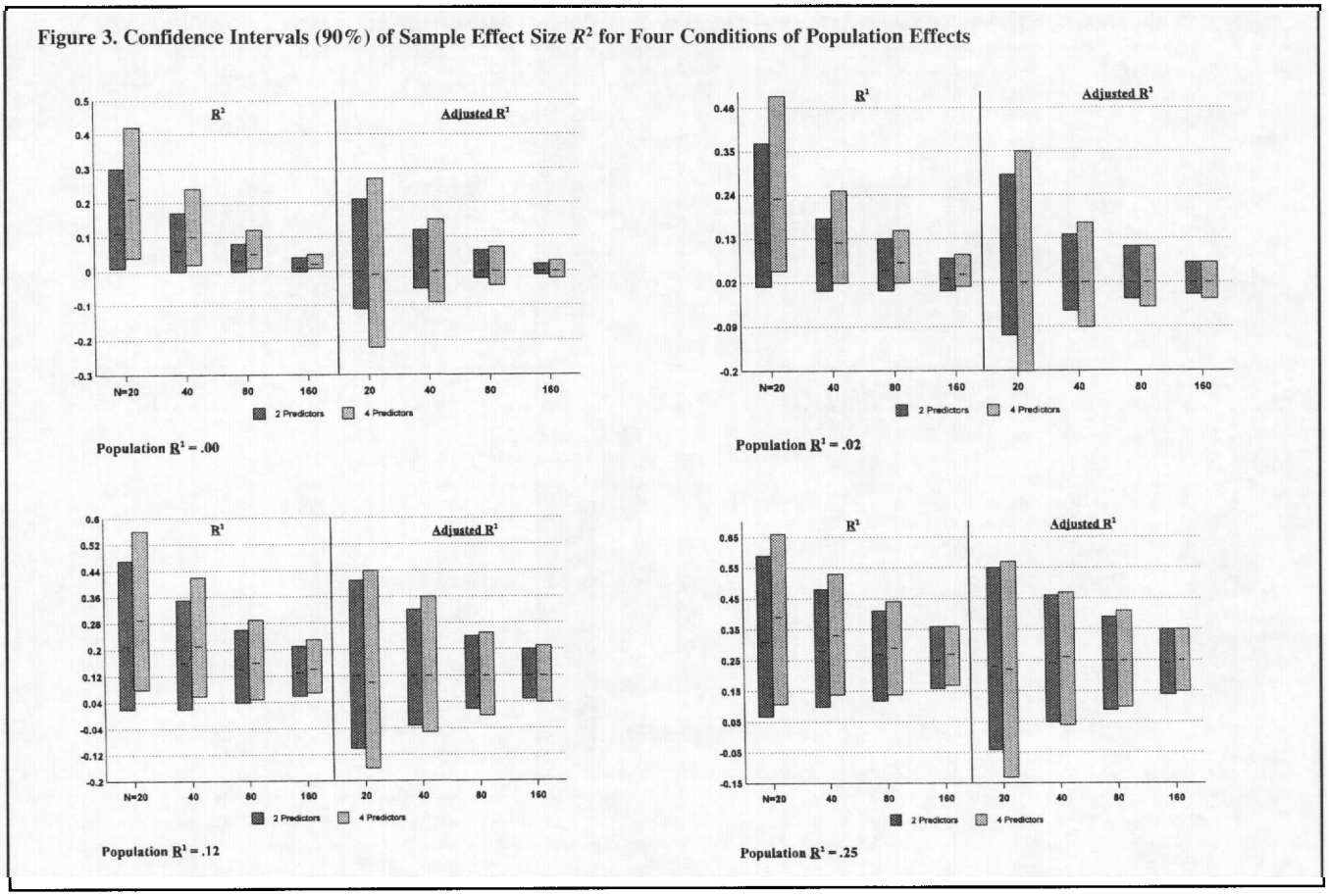
Second, there is considerable sampling variability of sample effect size d. For example, under the condition of population d = .00 (i.e., two samples drawn from the same population, thus no real difference between the two samples), for small-size condition such as $N = 20$ ($n_1 = n_2 = 10$), the 90% confidence interval almost covers the range from −.80 to +.80. In other words, for that sample-size condition, when two samples were drawn from the same population, and consequently, there was no real difference between the two groups, I could have obtained a large effect size (±.80) just by chance (i.e., sampling error). Even when sample size was increased to $N = 80$ ($n_1 = n_2 = 40$), probably a moderate sample size for many experimental designs, I still could have obtained sample effect size almost as large as ±.40 (moderate effect) by chance.

Third, the extent of sampling variability is obviously influenced by sample size. With the increase of sample size, the sampling variability of sample d, as represented by the 90% confidence intervals, shows a clear trend of becoming gradually smaller under all the conditions of population effect size (zero, small, medium, and large). That trend indicates that, if there are two identical effect sizes (e.g., moderate effect of d = .40) from two different studies involving different sample sizes (e.g., one is based on sample size of 40 [$n_1 = n_2 = 20$], and the other is based on $N = 160$ [$n_1 = n_2 = 80$]), the one based on the larger sample size is more trustworthy because such an effect size is very unlikely to have occurred because of sampling error. That result indicates that the use of effect-size measure should take sample size into consideration.

The sampling variability of another major type of effect-size measure is shown in Figure 3—the measure of association strength as represented by R^2. Because sample R^2 is widely known to have upward bias, I also reported one form of bias-corrected R^2 (adjusted R^2 obtainable from SAS or SPSS regression procedure) in Figure 3. The sampling variability of the R^2 and adjusted R^2 is represented by the 90% confidence interval bar; the mean R^2 based on 500 replications is represented by the short horizontal line within each confidence interval bar.

In addition to some common observations already discussed for the effect-size measure d in Figure 2, several observations unique for sample R^2 in Figure 3 can be made. First, whereas measure d in Figure 2 has been shown to be an unbiased estimator of population d, sample R^2 has obvi-

Figure 3. Confidence Intervals (90%) of Sample Effect Size R^2 for Four Conditions of Population Effects



ous upward bias, as indicated by the position of mean R^2 (short horizontal line within each 90% confidence interval) that is consistently, and sometimes considerably, above the population R^2 under all conditions. Bias correction, however, has worked well because the means of all sample adjusted R^2s are very close to the population R^2 value.

Second, sample R^2 from the four-predictor regression model has more upward bias than that from the two-predictor regression model. That finding is expected because under the same sample-size condition, the ratio of sample size to the number of predictors (N/p) is smaller for the four-predictor model than that for the two-predictor model. As is widely known in regression analysis, it is the ratio, rather than sample size per se, that largely determines the stability of regression analysis outcomes (Stevens, 1996; Yin & Fan, in press).

Both sample R^2 and adjusted R^2 show considerable sampling variability, which decreases as the sample size increases. The considerable sampling variability may make obtaining a medium and even large effect-size measure by chance relatively easy, even when the population effect size is zero or very small (R^2 = .02). For example, for population R^2 = .02 (very small effect) and for the four-predictor regression model, the upper 90% confidence limit of sample R^2 reaches as high as .46 (very large effect) for N = 20 and about .25 (large effect) for N = 40. That degree of sampling variability influenced by sample size (or N/p ratio) highlights the need for effect size to be considered in combination with sample size; used by itself, sample effect-size measure could be misleading.

Table 1 reports the percentages of statistically significant tests under different population effect-size and sample-size conditions. When the population effect is zero, approximately 5% of tests are statistically significant (underlined entries in the table), close to the specified nominal Type I error rate (α level). When population effect size is not zero, Table 1 entries represent the power of the statistical tests in rejecting the false null hypothesis. The tests can adequately detect the population effect (adequate statistical power is defined to be about .80 [Stevens, 1996]) only when the population effect is moderate to large (d = .50, .80) and the sample size is not small ($N \geq 40$; see boldface entries in Table 1).

Table 1.—Percentages of Statistically Significant Tests (α = .05)

Sample[a]	Population d .00	.20	.50	.80
20	*5.90*	7.35	18.05	37.25
40	*5.90*	8.85	32.75	65.30
80	*5.25*	14.00	54.40	**92.40**
160	*5.95*	22.50	**85.30**	**99.75**
240	*5.65*	32.80	**96.45**	**99.95**
Sample	**Population R^2 .00**	**.02**	**.12**	**.25**
20	*4.10*	7.60	21.00	47.30
40	*6.50*	9.50	44.40	**82.00**
80	*5.70*	18.70	**77.10**	**98.60**
160	*4.00*	31.20	**98.10**	**100.00**

Note. Tests can adequately detect population effect only when it is moderate to large (d = 0.50, 0.80) and sample size is not small ($N \geq 40$; see boldfaced numbers). When population effect is zero, about 5% of tests are statistically significant (see italicized numbers).
[a]For two group situations d, $N = n_1 + n_2$; $n_1 = n_2$.

One does not want to trust something that could have occurred by chance (Type I error); Table 1 shows that, when there is true effect, statistical tests may cause concern of Type II error. In other words, one may conclude that there is no effect when, in fact, there is. Balancing the two opposite logical errors requires the researcher (a) to understand the consequences of Type I and II errors, respectively; (b) to consider effect-size measure; and (c) to make decisions accordingly. Practical guidelines for combining statistical significance testing and effect-size measure in research practice are offered in Table 2. The content of Table 2 is self-explanatory; therefore, no explanation or discussion is required here.

Conclusions

In this article, I attempted to show that statistical significance testing and effect size are two related sides of the same coin; they complement each other but they do not substitute for each other. Good research practice requires that both sides should be taken into account to reach sound quantitative research decisions. The Monte Carlo experiment showed empirically that there is considerable variability of sample effect-size measure and that the extent of such variability is influenced by sample size. Because of the sampling variability of an effect-size measure, what appears to be practically meaningful effect size may be the result of sampling error, and, consequently, is not trustworthy. Statistical significance testing and effect-size measure serve different purposes; the sole reliance on either may be misleading. Some practical guidelines (see Table 2) are suggested for combining statistical significance testing and effect-size measure to make decisions in research practice.

NOTE

1. Because of a sampling error, a sample statistic (e.g., sample mean) will vary around population parameter (i.e., population mean). Over repeated sampling, if the average of a sample statistic is very close to the population parameter, the sample statistic is said to be an unbiased estimate of the population parameter. On the other hand, if the average of a sample statistic (e.g, sample R^2 discussed in the text) does not converge on the population parameter (either higher or lower) over repeated sampling, the sample statistic is said to be a biased estimate (either upward or downward bias) of the population parameter.

Table 2.—Guidelines for Combining Significance Test Outcome With Effect-Size Measure

	Effect size: Small	Effect size: Medium	Effect size: Large
No	1. It appears that there is neither statistical nor practical effect. 2. Unless future research indicates otherwise, null hypothesis is favored both statistically and practically.	1. Sample effect looks promising, but some caution is warranted in interpreting the effect size by itself because medium effect size could have been the result of chance, even if it may look practically meaningful. 2. If one is concerned about Type II error (there is true effect but one fails to find it), look closer at the power of the test because if the sample size is small, one may not have the statistical power to detect potential meaningful effect.	1. One has some evidence that meaningful effect exists, but a little caution is still warranted about this effect size because large effect size could have occurred by chance when sample size is small. 2. If one is concerned about Type II error, look critically at the lack of power of the statistical test. 3. Tentatively favor the practical significance of the effect, while keeping an open mind for further research findings.
Yes	1. Statistical significance is not accompanied by practical significance and could have been the result of statistical power (large *N*). 2. Considerable caution is warranted in interpreting the statistical significant findings, and they should not be interpreted to mean something practically meaningful.	1. It is very unlikely that the observed effect is due to statistical chance. 2. The magnitude of effect is practically meaningful in many areas of social and behavioral sciences. 3. Conclude that effect is meaningful both statistically and practically.	1. There is high degree of certainty that the observed effect is not due to chance statistically, and the magnitude of the effect is also practically meaningful. 2. Conclude with confidence that effect is meaningful both statistically and practically.

Note. The vertical and horizontal arrows indicate the increase in the degree of certainty for statistical and practical meaningfulness of the observed effect.

REFERENCES

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Carver, R. (1978). The case against statistical significance testing. *Harvard Education Review, 48,* 378–399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd. ed.). New York: Erlbaum.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98.

Fowler, R. J. (1985). Point estimate and confidence intervals in measures of association. *Psychological Bulletin, 98,* 160–165.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Education Researcher, 5,* 3–8.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston, MA: Allen & Bacon.

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

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179–182.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Education and Psychological Measurement, 56,* 746–759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *The Journal of Experimental Education, 61,* 378–382.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. *The Journal of Experimental Education, 61,* 383–387.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61,* 293–316.

Snyder, P., & Lawson, S. (1993). Effect size estimates. *The Journal of Experimental Education, 61,* 334–349.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Mahwah, 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. (1993). The use of statistical significance in research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61,* 361–377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Education Researcher, 25,* 26–30.

Yin, P., & Fan, X. (in press). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education.*

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