META-ANALYSIS AND THE PLANNING OF FUTURE STUDIES

James K. Brewer
Florida State University

ABSTRACT

A recent meta-analysis on ability grouping by Kulik and Kulik (1982) was used as an example to demonstrate how a researcher might plan for adequate sample sizes and power in future research. Data sufficient to estimate minimum sample sizes and power were gleaned from the meta-analysis study or obtained from its authors. Distributions of harmonic mean sample size and estimated power were displayed and ranges of estimates for each were presented using effect sizes from the meta-analysis with fixed α. Although effect sizes were relatively symmetric about .10, power and harmonic mean sample sizes were quite skewed. Recommendations for major professors, researchers and journal editors were made to assist in the evaluation and planning of research.

INTRODUCTION

The relatively recent epidemic of meta-analysis studies (see, for example, Andrews, et al., 1980; Kazrin, et al, 1979; Kulik, et al., 1980; Mabe and West, 1982; Resenthal and Rubin, 1978; and Smith and Glass, 1977) is providing the educational researcher with a much-needed quantitative and qualitative synthesis of research findings as well as a multitude of measures of effectiveness of treatments (Glass, 1976; Glass, McGraw and Smith 1981; Hedges, 1982; Hunter, Rosenthal and Rubin 1982; Kraemer and Andrews, 1982; and Schmidt and Jackson, 1982.) Two primary reasons for conducting and considering the results of meta-analyses are to provide researchers with information which may (1) help them to decide if further such studies are warranted and, if so, (2) assist them in the planning thereof. Consideration of the first reason is long overdue and, with or without meta-analyses, should have been initiated years ago to minimize the proliferation of studies reporting trivial and conflicting results. It is with the second reason, however, that this paper deals.

For decades educational researchers have planned and conducted studies using statistical methods with little or no consideration of the crucial ingredients of such methods. Notably, hypothesis tests have been conducted on every conceivable type and amount of data without an a priori incorporation or indication of what a meaningful treatment effect would be (effect size) or the probability with which this effect would be detected (power). Likewise, researchers typically omit any post hoc approximations of the magnitude of the treatment effect as found in the data. And only because tradition has such a firm grip on them have they bothered to mention the probability of a Type I Error (α) and then in an implicit form like, "$p < .05$".

It has been known since statistical inference began that one of the major elements of hypothesis testing, required sample size, is a function of α, power ($1-\beta$) and effect size (Neyman and Pearson, 1928). Presentday tables (see Cohen, 1969) and formulas (see Brewer, 1978; Davies, 1961; and Guenther, 1971) provide sample sizes for almost all commonly conducted parametric statistical tests. Why, then, have researchers not included these three in their plans for obtaining an adequate sample? The first reason is that behavioral statistics textbooks generally have neither discussed nor advocated using α, β and effect size in obtaining an adequate amount of data but have, rather, tacitly assumed the data were already collected. This "cart before the horse" mentality has denied researchers a readily available source for addressing one of their primary questions, "How much data do I need?". A second reason is that researchers knowledgeable of the philosophy of hypothesis testing have argued that only α could be present in their planning because power depended on a specific difference existing in the alternative hypothesis, H_a, and that this difference was never known. In short, what they have bemoaned is that they do not know the true effect size, therefore, they cannot set power and subsequently determine sample size. Although this argument may well be a cop-out from collecting a defensible amount of data, recent meta-analyses are providing, in quantitative form, information on both actual effect sizes and power which can be utilized by researchers in the planning of future studies, in particular, in obtaining an adequate amount of data. The purpose of this paper is to describe how power and effect size information may be extracted from an example meta-analysis study and utilized to determine adequate sample sizes for future studies.

An Example Meta-Analysis

In volume 19, No. 3, of the 1982 American Educational Research Journal, Kulik and Kulik published an article, "Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings." This informative study provided a nice vehicle for both describing effect sizes and power as well as illustrating their relationship to sample sizes. The authors were kind enough to provide a listing of the studies involved in their meta-analysis along with two-group sample size information, statistical t-test results and effect sizes for achievement test scores. Complete sample size, effect size, and test results were available for 42 of the 51 studies reported by Kulik and Kulik (1982) who also investigated several different ability grouping types and utilized several dependent variables. In this study no breakdown on ability grouping type was given and only the dependent variable, achievement, was considered. Even though approximately 40% of the studies in the meta-analysis were doctoral dissertations, these were not separately analyzed here or in the Kulik and Kulik (1982) study.

Effect Size

Effect size, the magnitude or degree of falsity of the null hypothesis when it is false, has been discussed in great detail for many different types of hypothesis tests (see, for example, Cohen, 1969) but still appears to be a stumbling block for researchers in attempting to find a minimum sample size (Brewer, 1978). For the two independent samples t-test of

$$H_o : \mu_1 - \mu_2 = 0$$

$$H_o : \mu_1 - \mu_2 \neq 0 \ ,$$

the alternate hypothesis, H_a, says there is a nonzero difference between μ_1 and μ_2. This true effect size is denoted δ and is never known. Cohen (1969) used ES to denote the true or population effect size as well as the magnitude which is meaningful or important to the researcher in an a priori context. Brewer (1978) and this writer distinguish between δ, the true effect size and ES, a meaningful effect size as judged by the researcher. Technically, one could denote yet another type of effect size, namely, an effect which is estimated from previously collected data, for example, from a meta-analysis. Rather than do this (and since estimated effect sizes from a meta-analysis could be meaningful to a researcher) ES will denote any nontrue, nonzero effect size either estimated from previous data or merely from the researcher's judgment of what a meaningful, important or worthwhile effect size would be if it existed. In the Kulik and Kulik (1982) meta-analysis ES was estimated for each specific statistical t-test by

$$ES = (\overline{X}_t - \overline{X}_c)/S_c \ , \qquad (1)$$

where $\bar{X}_t$, $\bar{X}_c$ and S_c are respectively the experimental (treatment) and control sample means and sample standard deviation for the control group. These ES values were then averaged over all 51 studies and found to be approximately .10. For the 42 studies in this present report the average ES value was .087, the median was about .14, and the modal value was .16. The frequency distribution of all 51 ES values is shown in the Kulik and Kulik (1982) paper, page 421, and was not duplicated here.

Converting this ES average to another measure of effect, the proportion of variance, using the point biserial correlation coefficient, r, (Cohen, 1969) gives

$$r^2 = (ES)^2/(ES^2 + 4) = .0019 \quad . \qquad (2)$$

This says that on the average, .2 of one percent of the achievement variability is accounted for by ability grouping. Effects of this magnitude could be referred to as "tiny" for want of a better word, but such magnitude judgments are immaterial for the present study.

Sample Size

The sample sizes in the 42 studies ranged from 17 to 522 per group where some samples were composite samples of several identical studies (see Kulik and Kulik, 1982). The harmonic mean per group sample size (which is essential for estimating the power of the tests) for experimental and control groups ranged from 23 to 430 with 20 of the 42 studies having equal sample sizes in the two groups. A display of the distribution of the harmonic means (rounded up to the nearest whole number) is shown in Table 1.

The mean of the harmonic means was 131 and the median was approximatey 96. The median is probably a better descriptor of the per group sample size due to the excessive skew of the sample size distribution. It may be of some interest to the reader to note that the Spearman's rho between sample sizes (n) and absolute value of effect size (ES) in the 42 studies was found to be -.058, showing virtually no rank order association between these two variables.

Table 1

Distribution of per group sample sizes
(Harmonic means)

n	f
1 - 50	12
51 - 100	10
101 - 150	5
151 - 200	5
201 - 250	4
251 - 300	3
301 - 350	0
351 - 400	1
401 - 450	2

Power

Power, in the usual two samples t-test, is the probability of correctly rejecting H_o, that is, of rejecting $H_o: \mu_1 - \mu_2 = 0$ given that $\mu_1 - \mu_2 = \delta$. When α, δ and n are known, power $(1-\beta)$ can be approximated from tables (see Cohen, 1969) or from equations derived from power curves (see Davies, 1961). Substantial criticism has been heaped on behavioral science research for having both very low power and failure to consider power in the planning of statistical reports. (Brewer, 1972; Chase and Tucker, 1975; Cohen, 1962; Haase, 1974). If educational researchers choose to conduct statistical tests with low power and justify this choice then the criticisms are unjustified, but if researchers believe they cannot control power, or at least estimate its minimum size, then meta-analyses may provide a partial response and solution to these justified criticisms.

A researcher who wished to compute the power from a meta-analysis could use the reported sample sizes and effect sizes and, either by utilizing power curves, published tables, or formulas, estimate the power. For example, in the Kulik and Kulik (1982) meta-analysis the two independent samples t-test was the statistical test which has an approximate sample size formula of the form (see Davies, 1961)

$$\sigma_1^2/n_1 + \sigma_2^2/n_2 = (\delta)^2/(Z_\alpha + Z_\beta)^2 \ , \qquad (3)$$

where Z_α and Z_β are respectively standard normal values such that

$P(Z \geq Z_\alpha) = {}^\alpha/2$ (for nondirectional tests)

or $P(Z \geq Z_\alpha) = \alpha$ (for directional tests),

$P(Z \geq Z_\alpha) = \beta$, when δ is the true difference

between μ_1 and μ_2, n_1 and n_2 are the two sample sizes and σ_1^2, σ_2^2 are the population variances for the two groups.

Assuming, as does the t-test, that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we find that

$$2(n_1)(n_2)/(n_1 + n_2) = 2\ [(Z_\alpha + Z_\beta)\sigma/\delta]^2 \qquad (4)$$

is the per group sample size for an independent samples t-test. The left side of (4) is simply the harmonic mean of n_1 and n_2. If δ is expressed as a function of the true standard deviation and approximated by the meta-analysis ES value given by (1), then equation (4) becomes

$$n = 2[(Z_\alpha + Z_\beta)/ES]^2 \ . \qquad (5)$$

Solving (5) for Z_β gives

$$Z_\beta = ES\sqrt{n/2} - Z_\alpha \ . \qquad (6)$$

For any ES, n and Z_α, Z_β can be found. Consulting a normal table will give power, since power = 1 - β, that is, $P\ (Z \leq Z_\beta)$. For many whole number harmonic mean values, n, the usual α levels and some ES values, Cohen's (1969) tables may be used with virtually identical results.

Equation (6) was utilized for all 42 studies in the Kulik and Kulik (1982) report since untabled ES values like .16 were found with n values such as 165. Table 2 displays the distribution of power values for the 42 studies using reported sample sizes, ES values and α = .05 (nondirectional).

Table 2

Distribution of power values from example meta-analysis (α = .05 (nondirectional))

Power	f
0 - .10	17
.11 - .20	4
.21 - .30	4
.31 - .40	5
.41 - .50	1
.51 - .60	2
.61 - .70	0
.71 - .80	3
.81 - .90	2
.91 - .99	4

If one used the average ES (.087) and n value (131) for the 42 studies along with α = .05 (nondirectional) then equation (6) yields an across-study power of .10. The average power of the 42 studies calculated separately was found to be .33, the median power was approximately .19 and the mode was .03. Due to the skewed distribution of power values, the median would most likely be a better central location descriptor of the power than the average.

Using Meta-Analyses Results

Suppose a researcher wished to conduct further studies in ability grouping and wanted to use the information provided by the Kulik and Kulik (1982) meta-analysis on the achievement variable. Assume further that the researcher was not interested in a particular subset of studies, but was willing to be guided by the results of all 42 studies herein discussed. (This may be an unwarranted assumption, but otherwise would involve a sub meta-analysis on the set of studies of interest).

Knowing that n, α , β and ES are intertwined and that setting any three determines the remaining value, the researcher must decide whether (1) to plan for an adequate sample given the effect size, α and power reported in the meta-analysis, (2) use the meta-analysis sample sizes, ES and α values to find the power for a future statistical test (3) or use some other combination of the four from the meta-analysis to find the remaining value. Probably the most reasonable approach would be to use the ES (or a range of ES values) from the meta-analysis, present α and β values which are satisfying to the researcher and calculate from this information an adequate sample size (or range of sizes). Following this approach and assuming, for example, that $\alpha = .05$, $\beta = .10$ (power = .90) and ES = .087 (from the meta-analysis) the researcher would find that the minimum per group sample size for a two-sample nondirectional t-test was, using equation (5),

$$\begin{aligned} n = n_1 = n_2 &= 2\ [(Z_\alpha + Z_\beta)/ES]^2 \\ &= 2\ [(1.96 + 1.28)/.087]^2 \\ &= 2774 \end{aligned}$$

This means a total sample size of 5548. Perceiving that this was an immodest minimal sample size requirement given the usual practical constraints of space, subject availability and cost, the thoughtful researcher could not help but wonder why so much data is required when the maximum per group sample used in the 42 meta-analysis studies was only 430.

The answer is reasonably simple and is not because of the plausible levels of .05 for α and .10 for β. It is because the average effect size is so very small. Even if the researcher used the modal ES value of .16 for the 42 studies, the minimal per group sample size would be around 800 for the same α and β. Adjusting α and β upward to unfamiliar levels of .20 each would still require a sample size of 352 per group when using the most commonly occurring ES value of .16. At some point the researcher may decide that further studies of ability grouping are too costly if the researcher wished to detect treatment effects similar to those central location values found in the Kulik and Kulik (1982) meta-analysis.

Taking a different tack, the researcher may decide to use the average sample size, α and ES, from the meta-analysis and determine what the power would be for such a study. An average ES of .087, the median per group sample size of 96, and $\alpha = .05$ (nondirectional) yields a power of .09 (from equation 6). This tells the researcher that if an infinite number of t-tests were conducted with this sample size, ES and α, the researcher could expect to correctly reject H_O about 9% of the time. (Note that the researcher would incorrectly reject 5% of the time due to an α of .05.) This kind of assurance of correctly rejecting an hypothesis is hardly worth writing home about. Even if the power were calculated for each of the 42 studies and the median of those values found, the power would still be only .19. (The modal power was even lower at .03 and using the average per group size of 131 gives a power of only .10.) For the researcher who believes that the average or median ES found in the

42 studies is a faithful and accurate reflection of what treatment differences are actually like and who wishes to keep sample sizes around the average or median of previous studies, the power prospects are indeed dismal. Modifying α drastically would provide only modest relief and a considerable increase in n might be too costly. For example, with α increased to .50 (nondirectional), ES at the average of .087 and n at the median value of 96, equation (6) gives a power of .43. To further illustrate the absurdity of extreme values, one need only set $\alpha = 1.0$ (nondirectional) for the same ES and n to produce a power of .73 which is still below the minimum power proposed by Cohen (1969). These last two choices for α , however, would hardly be reasonable for acceptable research.

The reader will note that, when a meaningful ES is chosen subjectively or estimated from a meta-analysis summary the value is, in a sense, a "threshold" value in that any true standardized mean differences smaller than this ES are considered trivial and any true differences larger than or equal to this ES are important. This makes whatever power and sample size calculations emanating from this ES (for fixed α) minimal. In order to approximate a reasonable range of power values, one could set (for fixed n and α) ES at the smallest of the central location values for ES from the meta-analysis to produce a lower limit for the power and set ES at the largest of the central location values to produce an upper limit for the power of the test. In the 42 studies of the Kulik and Kulik (1982) meta-analysis, the average ES of .087 was the smallest of the central location measures and the modal value of .16 was the largest. Likewise, using the largest of the central location measures for the per group sample size (the average n of 131) and the smallest central location measure for n (the median of 96), upper and lower limits for power could be derived for fixed α and ES.

Combining (with α = .05), for example, these upper and lower limits (larger n with larger ES and smaller n with smaller

ES) and substituting in equation (6) produces approximate upper (Up) and lower (Lp) power limits of

$$\begin{aligned} Up &= P(Z \leq ES\sqrt{n/2} - Z_{\alpha}) \\ &= P(Z \leq .16\sqrt{131/2} - 1.96) \\ &= P(Z \leq - .67) \\ &= .25 \end{aligned}$$

and

$$\begin{aligned} Lp &= P(Z \leq .087\sqrt{96/2} - 1.96) \\ &= P(Z \leq -1.36) \\ &= .09 \quad . \end{aligned}$$

These values provide the researcher with a plausible approximate range of expected power values given the extreme central location values of n and ES found in the meta-analysis (for α fixed at the usual .05 level, nondirectional).

Substituting the Lp value of .09 into equation (5) along with the largest central location ES value of .16 gives the lower limit (Ln) of the per group sample size (with α = .05) as

$$\begin{aligned} Ln &= 2[(1.96 - 1.34/.16]^2 \\ &= 30 \quad . \end{aligned}$$

The upper limit of the per group sample size, Un, will be found from equation (5) by substituting the smallest ES central location measure (the average of .087) along with the Up of .25 found above. Thus for α = .05,

$$\begin{aligned} Un &= 2[(1.96 - .25)/.087]^2 \\ &= 774 \quad . \end{aligned}$$

Clearly, there is no such thing as an "upper limit" to sample size except for financial or practical constraints, but these guides may provide some reasonable bounds, given the estimated limits of power along with extremes of ES for α fixed at .05. If, instead of the power bounds calculated above, one used the extremes of power and ES found in the Kulik and Kulik (1982) studies, that is, the modal level of .03 for power along with the modal ES of .16 (and the average power of .33 with the average ES of .087), the upper bound for n would be 602 and the lower bound for n would be 1. These limits are no more helpful than the values of 30 and 774 above, except that they result in an impossible per group sample size of 1 for a t-test. This latter situation is a result of the virtual zero value for the modal power of .03 found from the Kulik and Kulik (1982) report.

Discussion and Comments

Not all meta-analyses offer the researcher so little hope of using meta-analysis values to find an adequate sample or of having reasonable power when using the sample sizes from the meta-analysis. The very small effect sizes found in the Kulik and Kulik (1982) study may not, however, be atypical since Feldt (1972) reported finding quite small effect sizes in the standardized testing literature.

Regardless of the size of the estimated effect found in meta-analyses, the thorough researcher should consider them in preparing for or deciding to conduct further research in the area. Editors of journals could assist researchers in conducting meta-analyses and utilizing their results for planning purposes. If there were a question of sample size adequacy, for example, an editor could ask for a justification of the sample size(s). The justification could be in the form of published meta-analyses results or, in the absence of any meta-analyses, authors could provide another rationale for the choice of α, power and ES. In addition, editors should insist on a reporting of post hoc indicators of effect like eta squared, omega squared or some other measure (see Glass, McGraw and Smith, 1981, or Hedges, 1982), because today's post hoc effect measure could be tomorrow's a priori ES value if deemed meaningful by the researcher planning tomorrow's study. This not only will help plan future studies, but will address the crucial issue of practical importance so often missing from research reports. Authors of meta-analysis studies could also assist researchers in planning future research by providing not only effect size indicators, but α levels and sample sizes used in any statistical tests performed.

Although not publications in the usual research journal sense, doctoral dissertations comprised a large proportion (around 40%) of the Kulik and Kulik (1982) meta-analysis articles. Due to their relatively extensive reviews of the literature not found in most published reports, dissertations are a fertile field for the use of meta-analysis and it would appear to be imperative that dissertations contain some form of research synthesis, preferably one of the many forms referenced in this paper and elsewhere. As with any other meta-analysis, they should contain sufficient information on post hoc measures of effect, power and sample size for future researchers and dissertation writers to plan adequate studies.

For an editor or major professor to allow small sample size, but accept only results which have large post hoc effects (thereby keeping power high for modest α levels), although reasonable on the surface, is frought with difficulties. Large effects, for example, could result from small samples by virtue of a very small number of unobtrusively deviant observations and the research would be, in effect (no pun), capitalizing on outliers when there is no real effect present. The identification and resolution of outliers to minimize this problem is, in and of itself, no mean feat (See Fisher, 1980, and Woolley, 1981) and may be more difficult than obtaining larger sample sizes. In addition, what constitutes "a large effect" would probably not be generally agreed on by researchers and/or editors.

What should and could be agreed on, however, is that all reported inferential studies, regardless of their statistical outcome, contain information on a priori effect size, α , β and n as well as some post hoc measures of effect. This way the reader will know what planning went into the determination of sample size and what treatment effect was indicated from the data.

REFERENCES

Andrews, G., Guiter, B. and Howie, P. Meta-analysis of the effects of stuttering treatment. Journal of Speech and Hearing Disorders, (in press).

Brewer, J.K. On the power of statistical tests in the American Educational Research Journal. AERJ, 1972, 9, 391-401.

Brewer, J.K. Effect Size: The most troublesome of the Hypothesis Testing Consideration. Center on Evaluation, Development and Research Quarterly, 1978, 2, No 4.

Chase, L.H. and Tucker, R.K. A power-analytic examination of contemporary communications research. Speech Monographs, 1975, 42, 29-41.

Cohen, J. The statistical power of abnormal social psychological research. Journal of Abnormal and Social Psychology, 1962, 65, 145-153.

Cohen, J. Statistical power analyses for the behavioral science. New York: Academic Press, 1969.

Davies, O.L. Statistical Methods in Research and Production (Chapter 5). New York: Hafner Publishing Co., 1961.

Feldt, L.S. What Size Sample for Methods/Materials Experiments? Journal of Educational Measurement, 1973, 10, No. 3, 221-226.

Fisher, P.L. An investigation of outlier definition and the impact of the masking phenomenon on several statistical outlier tests. Unpublished doctoral dissertation, Florida State University, 1980.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G.V., McGaw, B. and Smith, M.L. Meta-analysis in Social Research. Beverly Hills: Sage Publications, 1981.

Guenther, William C. Sample Size Formulas for Normal Theory T-tests. The American Statistician, 1981, 35, 243-244.

Haase, R.F. Power analysis of research in counselor education. Counselor Education and Supervision, 1974, 14, 124-132.

Hedges, L.V. Estimation of Effect Size From a Series of Independent Experiments. Psychological Bulletin, 1982, 92, 490-499.

Hunter, J.E., Schmidt, F.L. and Johnson, G.B. Integrating Research Findings Across Studies. Beverly Hills: Sage Publications, in press.

Kulik, J.A., Cohen, P.A. and Ebeling, B.J. Effectiveness of programmed instruction in higher education: A meta-analysis of findings. Educational Evaluation and Policy Analysis, 1980, 2, 51-64.

Kulik, C.L. and Kulik, J.A. Effects of Ability Grouping on Secondary School Students: A Meta-analysis of Evaluation Findings. American Educational Research Journal, 1982, 19, No. 3, 415-428.

Mabe, P.A. and West, Stephen. Validity of Self-Evaluation of Ability: A review and meta-analysis. Journal of Applied Psy, 1982, 67, No. 3, 280-296.

Neyman, J. and Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika, 1928, 20A, 175-263.

Rosenthal, R. and Rubin, D.B. Interpersonal expectancy effects: the first 345 studies. The Behavioral and Brain Sciences, 1978, 3, 377-415.

Rosenthal, R. and Rubin, D.B. Comparing effect size of independent studies. Psychological Bulletin, 1982, 92, #2, 500-504.

Smith, M.L. and Glass, G.V. Meta-analysis of psychotherapy outcome studies. American Psychologist, 1977, 32, 752-760.

Woolley, Tom. An investigation of the effect of the swamping phenomenon on several block procedures for multiple outliers in univariate samples. Unpublished doctoral dissertation, Florida State University, 1981.