Going Beyond the Literature Review with Meta-analysis

Keith McNeil, New Mexico State University
Isadore Newman, The University of Akron

Abstract

Meta-analysis is a procedure that transforms research results into a common metric--called the effect size. This effect size can be aggregated if consistent across studies. If the effect size is not consistent, study characteristics can be used to ascertain why the effects are not consistent. The focus of this paper is on encouraging linear, curvilinear, and interactive investigations of the relationship between study characteristics and effect size.

In a previous paper (McNeil & Newman, 1994), we reviewed how one can obtain an Effect Size in order to aggregate the results of several similar studies. (See Glass, McGaw, & Smith, 1981; Light & Pillemer, 1984; and Rosenthal, 1984 for more details.) If the Effect Sizes are fairly similar, they can be aggregated to produce an average effect size. In many instances, Effect Sizes will vary. These discrepant results may be due to problems with internal validity, problems with external validity, or to random errors. We present a number of situations in which the researcher can uncover the reasons for the discrepant results. We rely on the General Linear Model to do the detective work to uncover the reason(s), because of its wide applicability.

Similarly labeled treatments or participants may differ in important ways

Although a researcher may refer to a set of treatments as similar, the researcher may find that the treatments differ in terms of some attribute, as depicted by Class Size in Table 1. Although the total mean Effect Size in Table 1 is .418, one could easily test the difference between the small Class Size studies (mean of .488) and the large Class Size studies (mean of .348), producing a *t* value of 1.85 with an associated directional probability of .051. These results (given an alpha of .05) would not lead the researcher to conclude that the treatment (however defined and however tested) is more effective with small classes than with large classes.

The above test could be accomplished with a *t*-test of the difference between two independent means, or with the comparison of two regression models. The research hypothesis in this case is: "Small classes produce larger Effect Sizes than do larger classes." The criterion is Effect Size, and the information known about the subjects is whether the results come from large classes or small classes. Therefore the Full Model, containing the full amount of information, is: Effect Size = a*U + b*S/L + E1 (where S/L = 1, if small Class Size, 0 if large Class Size).

Since this is a directional research hypothesis, we want the weighting coefficient to be greater than 0, which means that the restriction on the Full Model is: b = 0, resulting in the following Restricted Model: Effect Size = a*U + E2. The R^2 of the Full Model (itself an effect size of large versus small class size) is .30. When compared to the R^2 of 0.00 of the Restricted Model, this results in a p value of .051, the same as that for the *t*-test. (See McNeil, Kelly, & McNeil, 1975 or McNeil, Newman, & Kelly, *in press*, for extended discussion of testing research hypotheses with the General Linear Model.) Using an alpha of .05, we would not have obtained significance.

Table 1

A meta-analysis investigating class size differences in effect size

Study	Δ	Class Size	Small (S) or Large (L) Class Size
1	.35	36	L
2	.45	25	S
3	.60	15	S
4	.40	30	S
5	.70	8	S
6	.30	40	L
7	.31	45	L
8	.29	35	S
9	.40	40	L
10	.38	43	L

Mean Effect for all studies: = .418
Mean Effect for studies with large classes: = .348
Mean Effect for studies with small classes: = .488

One also could look at Class Size as a continuous variable rather than as an artificially dichotomized variable. The data in Table 1 have been plotted in Figure 1, treating Class Size as a continuous variable. The linear correlation between Class Size and Effect Size yields a correlation of .91 and an R^2 of .83. While the difference in Effect Size between large and small classes was not significant, the correlation between Class Size and Effect Size is significant.

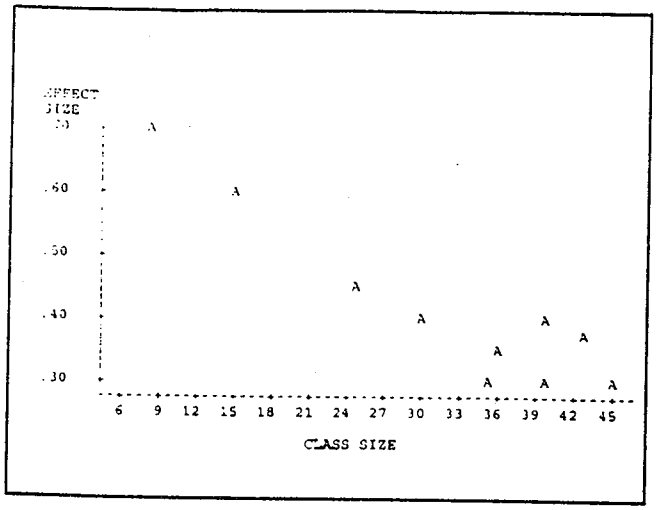

Figure 1. Relationship between Class Size as a continuous variable and Effect Size. Fictitious data from Tables 1 and 3.

The above correlation can be obtained and tested for significance with GLM. The research hypothesis is: "There is a negative linear relationship between Class Size and Effect Size," resulting in the following Full Model: Effect Size = a*U + b*Class Size + E1 (where Class Size is a continuous variable). Since there is expected to be a negative relationship, the expectation is that the weight, b, is less than 0, resulting in the restriction on the Full Model of b=0. When the restriction is placed on the Full Model, the following Restricted Model obtains: Effect Size = a*U + E2. Comparing the two models results in a p value of .0002, less than the *a priori* alpha of .05. Thus there is a significant negative relationship between Class Size and Effect Size.

Early discussions of effect size focused on differences between groups and linear relationships. Rosenthal (1980) and Light and Pillemer (1984) emphasized plotting the data to look for non-linear and interacting relationships in the data. The advantage of the General Linear Model is that non-linear and interactive relationships can easily be tested empirically. Inspection of Figure 1 supports the investigation of a second-degree curve.

The above assertions were based on the testing of the following research hypothesis: "There is a second-degree relationship between Class Size and Effect Size, over and above the linear fit." In order to allow for a second-degree curve, a second-degree component of Class Size must be added to the previous straight-line model, resulting in the Full Model: Effect Size = a*U + b*Class Size + c*Class Size^2 + E3, where Class Size^2 is simply the squared value of Class Size. Since the research hypothesis specifies a non-directional second-degree relationship, the expectation is that the weight for the second-degree component be not equal to 0, resulting in the restriction that the weight is equal to 0. If the weight, c, is forced to be equal to 0, the Full Model becomes the following Restricted Model: Effect Size = a*U + b*Class Size + E4. Comparing the R^2 of the two models (.91 and .83) results in a p value of .036, indicating a significant second-degree relationship.

Treatments may be more or less effective depending upon who the subjects are, the setting in which the treatment occurs, or other situational variables

An aptitude-by-treatment interaction is often common in educational research, and can provide valuable insight to a discipline (Cronbach & Snow, 1977; Tobias, 1976). One should not expect a given treatment to work equally well for all types of subjects. Therefore, one should not aggregate results from studies that produce different results, as those in Table 2 do.

Table 2

A meta-analysis identifying class size differences interacting with major on effect size

Study	Δ	Class Size	Major (M) or Non-major (N)
1	.10	5	N
2	.10	10	N
3	.10	15	N
4	.20	20	N
5	.30	30	N
6	.50	50	N
7	.75	75	N
8	.80	80	N
9	.80	20	M
10	.75	25	M
11	.50	50	M
12	.70	30	M
13	.25	75	M
14	.20	80	M
15	.15	80	M
16	.25	80	M

Mean Effect for all studies: = .403
Mean Effect for studies with majors: = .450
Mean Effect for studies with non-majors: = .356

Testing the research hypothesis: "Majors produce a larger Effect Size than do Non-Majors" would require the same analysis as in Table 1. The Full Model would be: Effect Size = a*U + b*M/M + E1 (where M/M = 1 if Major, 0 if Non-major). The Restricted Model would be: Effect Size = a*U + E2. The R^2 of the Full Model is .03, resulting in a p value of .5158--no significant difference between major and non-major.

It may be that the apparent inconsistency in results is due to the nature of the treatment (instruction). While there is no overall difference between Majors and Non-majors, the results are clearer when the interaction between whether the course is restricted to Majors and Size of the classroom is considered, as indicated in Figure 2. Why this is the case is not known at this time, but one possibility is that learning requires some content literacy, and content

literacy is facilitated by small classes and debilitated by large classes. This finding would help the researcher in identifying a moderator variable which could be tested in future research.

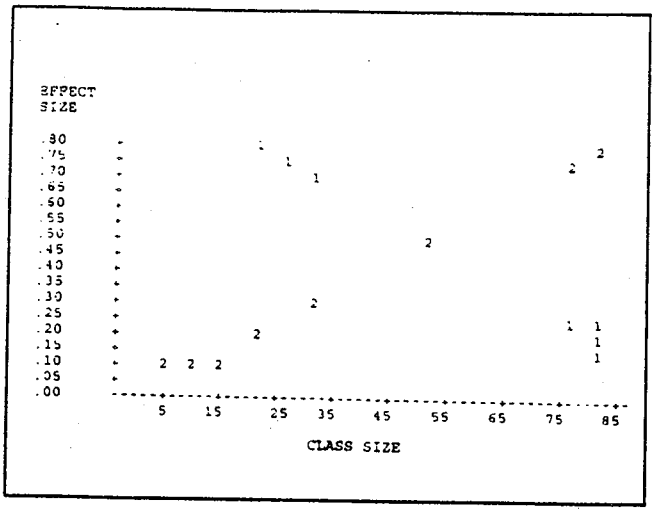

Figure 2. Relationship between Class Size as a continuous variable and Effect Size for Majors (1) and Non-majors (2), using Fictitious data from Table 2.

The above discussion relates to testing the research hypothesis: "There is an interaction between Class Size and Major/Non-major in the prediction of Effect Size." The Full Model needed to reflect all the information in the research hypothesis is: Effect Size = a*U + b*Class Size + c*M/M + d*(Class Size*M/M) + E1 (where Class Size*M/M is simply the product of Class Size and M/M). If there is the expected interaction, then the weight, d, will be non-zero. If there is no interaction, then the weight, d, will be zero, resulting in the Restricted Model: Effect Size = a*U + b*Class Size + c*M/M + E2. The R^2 of the Full Model is .99, while the R^2 of the Restricted Model is .03. The *F*-test of these two models results in a p value of .0001, and since this value is less than our alpha of .05, we can conclude that there is an interaction between Class Size and Major/Non-major.

The type of research design employed in a study can strongly influence the outcome

Different results can occur as a function of how the researcher designed the study. For instance, studies in most research areas document that Volunteers react differently than Non-volunteers. Rosenthal and Rosnow (1975) reviewed the research on the differences between Volunteers and Non-volunteers. Some of the differences are that Volunteers tend to (a) be better educated, (b) have higher social class, (c) be more intelligent, and (d) have a higher need for social approval. Suppose that the studies in Table 1 were conducted with college sophomores, half were conducted with Volunteers while the other half were conducted with Non-volunteers (Non-volunteer subjects were randomly assigned to experimental groups as part of their course requirements), as indicated in Table 3.

Table 3

A meta-analysis identifying volunteer differences in Effect Size

Study	Effect Size	Class Size	Volunuteer (V) or Non-Volunteer (N)
1	.35	36	N
2	.45	25	V
3	.60	15	V
4	.40	30	V
5	.70	8	V
6	.30	40	N
7	.31	45	N
8	.29	35	N
9	.40	40	V
10	.38	43	N

Mean Effect for all studies: = .418
Mean Effect for studies with non-volunteers: = .326
Mean Effect for studies with volunteers: = .510

Now the apparent consistency in the results is due to whether the subjects volunteered for the study. Indeed, the Volunteer Effect is (.510 - .326) = .184. Why this is so can only be conjectured at this time, although there is much in the research design literature about demand characteristics. Volunteers usually want the researcher to succeed, are extremely willing to do whatever requested, try to figure out what the researcher wants to do, attend to cues diligently, etc.

The research hypothesis tested here would be of the same structure as the ones in Table 1 and Table 2: "Volunteers produce a larger Effect Size than do Non-volunteers." This research hypothesis results in the following Full Model: Effect Size = a*U + b*V/NV + E1. Since this is a directional Research Hypothesis, we want b to be greater than 0. The Full Model R^2 is .51, and the Restricted Model R^2 is 0.00, resulting in a p value of .018. For these fabricated data, Volunteers produce a larger effect size than do Non-volunteers.

The particular analysis procedure that is used may be related to outcomes

One of the continuing concerns is the unit of analysis problem. For instance, should an educational researcher use the individual subject as the unit of analysis, or should the classroom mean be used as the unit of analysis? If the teacher effect is potent, then using the classroom as the unit of analysis makes sense since all of those students in the one classroom were taught by the same teacher. If the treatment and dependent variable can be influenced by the entire school--a school effect--then it makes sense to use the school as the unit of analysis. Which level a researcher uses can influence heavily the magnitude of Effect Size. Pillemer and Light (1980) stated that the more highly aggregated the unit of analysis, the stronger the relationship will be.

Proposed solution when several results from one study are analyzed

One study might contribute more than one Effect Size in any one meta-analysis. This could occur if multiple dependent variables were used, if multiple populations were investigated, or if multiple treatments or multiple comparison groups were used. In these cases, the unique aspects of the study impinge, to some extent, on each of the Effect Sizes collected from that one study. To avoid the problems of non-independent data, one can extend the analysis of repeated measures to such study-results (Tracz, Newman, & McNeil, 1986).

Suppose the 10 Effect Sizes in Table 3 actually came from six different studies. The analysis would need to take into consideration the fact that there is non-independence in the data--some of the studies supplied more than one Effect Size. The proposed solution is to include "study vectors," analogous to "person vectors" in repeated measures analysis. The study vectors are presented in Table 4. The research hypothesis tested is "Small classes produce larger Effect Sizes than do Larger classes, over and above the individual differences due to each study." The Full Model would need to have not only information about size of class (S/L), but also which of the six studies the results were from (S1, S2, etc.). We thus have as the Full Model: Effect Size = a*U + b*S/L + c*S1 + d*S2 + e*S3 + f*S6 + g*S8 + h*S9 + E1 (where S1 = 1 if Effect Size is from study 1, 0 otherwise, etc.). If there is no difference between large and small Class Sizes (over and above study differences), then the weight, b, will be equal to 0, resulting in the Restricted Model: Effect Size = a*U + c*S1 + d*S2 + e*S3 + f*S6 + g*S8 + h*S9 + E2. The R^2 of the Full Model is .71 and the R^2 of the Restricted Model is .67, resulting in a p value of .4929--small classes do not produce larger Effect Sizes than do Larger classes, over and above study differences.

Table 4

A meta-analysis investigating class size differences in effect size, considering multiple results from several studies

Study	Δ	Small (S) Large (L)	S1	S2	S3	S6	S8	S9
1	.35	L	1	0	0	0	0	0
2	.45	S	0	1	0	0	0	0
3	.60	S	0	0	1	0	0	0
3	.40	S	0	0	1	0	0	0
3	.70	S	0	0	1	0	0	0
6	.30	L	0	0	0	1	0	0
6	.31	L	0	0	0	1	0	0
8	.29	S	0	0	0	0	1	0
9	.40	L	0	0	0	0	0	1
9	.38	L	0	0	0	0	0	1

Mean Effect for all studies: = .418
Mean Effect for studies with large classes: = .348
Mean Effect for studies with small classes: = .488

Summary

Since the study characteristic is constant for any one study, none of the original researchers could have tested any of the research hypotheses discussed in this paper. Analyzing Effect Size facilitates the explanation for why different Effect Sizes are obtained from different studies. Such information is invaluable for understanding and extending the knowledge base in any field.

References

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Irvington.

Glass, G., McGaw, B., & Smith, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

McNeil, K., Kelly, F. J., & McNeil, j. T. (1975). *Testing research hypotheses using multiple linear regression*. Carbondale, IL: Southern Illinois University Press.

McNeil, K. Newman, I., & Kelly, F. J. (In Press). *Testing research hypotheses with the general linear model*. Carbondale, IL: Southern Illinois University Press.

McNeil, K. & Newman, I. (1994). Summarizing the literature review with meta-analysis. *Mid-Western Educational Researcher, 7*(3), 26-29.

Pillemer, D. B., & Light, R. J. (1980). Benefiting from variation in study outcomes. In R. Rosenthal (Ed.), *Quantitative assessment of research domains* (pp. 1-12). San Francisco: Josey-Bass.

Rosenthal, R. (1980). Combining probabilities and the file drawer problem. *Evaluation in Education: An International Review Series, 4*, 18-21.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: John Wiley.

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research*, 46, 61-74.

Tracz, S., Newman, I., & McNeil, K. (1986, April). Tests of dependence in meta-analysis using multiple linear regression. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.