Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal.

Volume 10 Number 14, September 2005 ISSN 1531-7714

Using the Binomial Effect Size Display (BESD) to Present the Magnitude of Effect Sizes to the Evaluation Audience

Justus J. Randolph, University of Joensuu, Finland &
R. Shawn Edmondson, Utah State University

The use of Rosenthal's Binomial Effect Size Display (BESD) as a tool for reporting the magnitude of effect sizes to the evaluation audience is discussed. The authors give an overview of the BESD, describe how it is calculated, and present a review of its strengths and weaknesses. Additionally, suggestions for the appropriate use of the BESD are given. An effect size to BESD conversion table is included.

Effectively communicating evaluation information to clients is a critical part of the evaluation process, yet many evaluators fail to give it careful consideration (Worthen, Sanders, & Fitzpatrick, 1997). The value of evaluation information, regardless of how scientifically defensible it is, can be undermined when it is not presented in language which is practical and easily understood by the intended audience. Weiss and Bucuvalas (as cited in Worthen et al., 1997) suggest that there are two important types of values to consider to ensure that evaluation information is useful to policy makers:

> *Truth value* refers to the technical quality of the study and to whether the findings correspond to policy makers' previous understanding and experience with how the world works (expectations) . . . *Utility value* refers to the extent to which the study provides explicit and practical direction on matters the policy makers can do something about and challenges the status quo (with new formulations and approaches). (p. 410)

One way that evaluators can increase the utility value of evaluation information is to make the statistics that they report more meaningful to the evaluation audience.

May (2004) provides three guidelines for making statistics more meaningful:

1. Understandability: the results should be reported in a form that is easily understood by most people by making minimal assumptions about the statistical knowledge of the audience and avoiding statistical jargon.
2. Interpretability: a statistic is interpretable when the metric or unit of measure that it is based upon is familiar or easily explained.
3. Comparability: the reported sizes of the statistics that might be compared can be

> compared directly, without any further manipulation. (pp. 527-528)

The *r*-based binomial effect size display (BESD) is particularly strong in terms of understandability and, to a lesser extent, interpretability. Because of the understandability and interpretability of the BESD, we argue that it can help improve the utility value of evaluation. In the next section of this article we describe the BESD and its calculations. We then give an overview of its benefits and disadvantages. We end with our suggestions for presenting the BESD to the evaluation audience.

THE BINOMIAL EFFECT SIZE DISPLAY

Although the coefficient of determination (r^2) is often used as an effect size that describes the proportion of variance accounted for in a dependent variable by predictor variables, this technique is problematic. Squaring *r* can make a practically significant effect appear to be insignificant, especially to a lay audience. The BESD is a tool that may be used to display the practical importance of an effect without relying solely on *r* or r^2 values or other less intuitive effect size measures (Rosenthal & Rubin, 1982). (In this article, by *BESD* we refer to only the *r*-based BESD and not to raw-data two-by-two tables. See May [2004] for a discussion on making raw-data two-by-two tables more meaningful.)

The BESD illustrates the practical importance of an effect by displaying a point-biserial *r* as a two-by-two contingency table. It presents the correlation simply and intuitively as the difference in outcome rates between experimental and control groups. The rows in a BESD table display the independent variable as a dichotomous predictor, such as belonging to an experimental or control group. The columns in the table display the dependent variable as a dichotomous outcome, such as improved and not improved. The row and column totals always add up to 100. Table 1 presents several different examples of BESDs along with their associated correlations.

As Table 1 illustrates, the effect size of a meta-analysis of psychotherapy interventions was reported to be $r = .32$. The BESD shows a psychotherapy success rate of 66% and a control group success rate of 34%. As Rosenthal, Rosnow, and Rubin (2000) state, " . . . an *r* of .32 (or an r^2 of .10) will amount to a difference between rates of improvement of 34% and 66% if half the population received psychotherapy and half did not, and if half the population improved and half did not" (p.17). If psychotherapy had no effect, then each cell in the BESD table would have been 50%.

The values in the BESD should be interpreted as "standardized" percentages, where the percentages within the cells have been set so that all margins are equal. By adjusting the percentages in each of the four cells of a two-by-two table so that row and column margins are equal, the BESD maximizes the symmetry of the cells. It is also important to recognize that the BESD assumes a 50% base-rate for both the experimental and control groups - an artificial situation created to illustrate the impact of the effect.

CALCULATING THE BESD

Creating a BESD for two groups with equal n-size and with homogenous variances is straightforward. To calculate the success rate of the treatment group, the formula, $(.50 + r/2)$, is used. To calculate the success rate of the control group the formula, $(.50 - r/2)$, is used. For example, if the value of *r* is 0.07, as in the Vietnam service and alcohol use example in Table 1, then the success rate of the treatment group is $(.50 + 0.07/2) = 53.5\%$. The success rate of the control group is $(.50 - 0.07/2) = 46.5\%$. Taking into consideration how cells A, B, C, and D are positioned in Table 2, putting the treatment group success rates into cells A and B and the control group success rates into cells C and D creates a BESD table identical to the Vietnam and alcohol use example in Table 1.

Table 1: *Examples of Binomial Effect Size Displays*

Measure	Variable		Total
Vietnam service and alcohol problems (r = .07)			
	Problem	No problem	
Vietnam veteran	53.5	46.5	100
Non-Vietnam veteran	46.5	53.5	100
Total	100.0	100.0	200
AZT in the treatment of AIDS (r = .23)			
	Death	Survival	
AZT	38.5	61.5	100
Placebo	61.5	38.5	100
Total	100.0	100.0	200
Benefits of psychotherapy (r = .32)[a]			
	Less benefit	Greater benefit	
Psychotherapy	34.0	66.0	100
Control	66.0	34.0	100
Total	100.0	100.0	200

Note. AZT = aziothymidine, AIDS = acquired immune deficiency syndrome. From "How are we doing in soft psychology?" by R. Rosenthal, 1990, *American Psychologist,* 50, p. 776. Copyright 1990 by the American Psychological Association. Reprinted with permission.

[a]The analogous r for 345 studies of interpersonal expectancy effects was essentially the same (Rosenthal & Rubin, 1978).

Below, Strahan (1991) explains how the BESD formulas presented in the preceding paragraph are derived from the phi coefficient formula:

> By assuming an equal-marginals contingency table, one can solve for any one – hence for all four – of the cell frequencies [in a BESD display] by working backward from the phi coefficient formula. Specifically, setting a + b = c + d = a + c = b + d = 100, it follows algebraically that d = a and b = c = 100 – a, so that from phi = r = (ad – bc)/[(a + b) (c + d) (a + c) (b + d)]$^{1/2}$, one gets r = a/50 – 1, and a = 50 + 50r. . . . From this it follows that the treatment success rate is .50 + r/2, and the control success rate is .50 – r/2. (p. 1084)

Table 2: *BESD Template*

Group	Improved	Didn't Improve	Total
Treatment	(A)	(B)	100
Control	(C)	(D)	100
Total	100	100	200

A BESD table can also be calculated from a standardized mean difference effect size (Cohen's d) using the formula, $r = d/\sqrt{(d^2 + 4)}$, when there are two groups with equal n-size. The BESD table can then be calculated from r using the formulas given above. Rosenthal et al. (2000) is an excellent resource for calculating BESD values for unequal n-sizes or for experiments involving more than two groups. Table 3 gives a list of Cohen's d values (from 0 to 3 in intervals 0.10), their associated r

values, and BESD values for two groups with equal n-size.

Table 3: *Effect Size to BESD Conversions*

Cohen's *D*	*r*	BESD Cells			
		A	B	C	D
0.00	0.00	50.0	50.0	50.0	50.0
0.10	0.05	52.5	47.5	47.5	52.5
0.20	0.10	55.0	45.0	45.0	55.0
0.30	0.15	57.4	42.6	42.6	57.4
0.40	0.20	59.8	40.2	40.2	59.8
0.50	0.24	62.1	37.9	37.9	62.1
0.60	0.29	64.4	35.6	35.6	64.4
0.70	0.33	66.5	33.5	33.5	66.5
0.80	0.37	68.6	31.4	31.4	68.6
0.90	0.41	70.5	29.5	29.5	70.5
1.00	0.45	72.4	27.6	27.6	72.4
1.10	0.48	74.1	25.9	25.9	74.1
1.20	0.51	75.7	24.3	24.3	75.7
1.30	0.54	77.2	22.8	22.8	77.2
1.40	0.57	78.7	21.3	21.3	78.7
1.50	0.60	80.0	20.0	20.0	80.0
1.60	0.62	81.2	18.8	18.8	81.2
1.70	0.65	82.4	17.6	17.6	82.4
1.80	0.67	83.4	16.6	16.6	83.4
1.90	0.69	84.4	15.6	15.6	84.4
2.00	0.71	85.4	14.6	14.6	85.4
2.10	0.72	86.2	13.8	13.8	86.2
2.20	0.74	87.0	13.0	13.0	87.0
2.30	0.75	87.7	12.3	12.3	87.7
2.40	0.77	88.4	11.6	11.6	88.4
2.50	0.78	89.0	11.0	11.0	89.0
2.60	0.79	89.6	10.4	10.4	89.6
2.70	0.80	90.2	9.8	9.8	90.2
2.80	0.81	90.7	9.3	9.3	90.7
2.90	0.82	91.2	8.8	8.8	91.2
3.00	0.83	91.6	8.4	8.4	91.6

STRENGTHS OF THE BESD

Under certain conditions, the BESD has several strengths that make it desirable for reporting the practical significance of effect sizes to lay audiences. Its strengths are listed below:

- The BESD is intuitively understood by lay audiences compared to somewhat complicated statistics such as *r*, r^2, or *d* (Rosenthal, 1990). That is, it has more of what May (2004) calls understandability and interpretability than *r*, r^2, or *d*.
- It is easy to compute.
- It is appropriate for understanding r^2 in the context of, as Rosenthal et al. calls them, "the 'softer, wilder' areas of the social and behavioral sciences – where the results often seem ephemeral and unreplicable, and where r^2 seems always to be too small" (2000, p. 25).
- It allows an evaluator to present a two-by-two table to evaluation audiences when there is not enough information to construct a two-by-two table from raw data. This might be the case when it is necessary to illustrate the magnitude of a combined, meta-analytic effect size to an audience who would understand and interpret a two-by-two display much better than a statistically-loaded effect size, like *r*, r^2, *d,* or an odds ratio.

CRITICISMS OF THE BESD

Researchers have criticized the BESD for a number of reasons (Crow, 1991; Strahan; 1991, McGraw, 1991; and Thompson & Schumaker, 1997). Some critics argue that the BESD is a misleading 'what-if' technique. Other critics argue that Rosenthal's BESD distorts results as a function of the symmetry of the raw data of the cells. These criticisms are explained in more detail below.

First, although the BESD may be intuitive for lay audiences, McGraw argues that "creating an artificial case that is correlationally equivalent to the original case so distorts the original data that the exercise is terribly misleading" (1991, p. 1084). For example, the lay audience may mistake the 'standardized' percentages of the BESD for the actual raw data if the BESD is not carefully explained. What's more, Strahan (1991) calls the BESD a "what if?" statistical technique, such as analysis of covariance, and therefore has all the faults of "what if?" techniques.

Second, Thompson and Schumacker (1997) make the case that as the asymmetry of the BESD

increases (i.e., the cells diverge further from 50%) the interpretation of the effect becomes more erroneous. For example, when the binary success rate is symmetrical at 50%, the percentage difference between Φ and *d* is 0 (Φ is the Phi coefficient - the measure of the degree of association between two binary variables.) However, when the binary success rate is 100%, the percentage difference between Φ and *d* is 100.

Rosenthal (1991) responds to these criticisms by saying that there are instances where the Pearson product moment correlation and its equivalents (i.e., the BESD) are not the proper effect sizes to report. Rosenthal concedes that in cases where asymmetry is very pronounced, the relative risk index or the difference in raw proportions are the most appropriate estimators of effect size. However, when holding the value of *r* constant but changing the symmetry of the cells, Rosenthal has shown that the differences of percentages between cases of minimum symmetry and maximum symmetry (when the BESD is used), vary only slightly. Other measures of effect size like relative risk or odds ratios vary considerably when the value or *r* for a two-by-two table is held constant but the symmetry of the cells is changed (see Rosenthal, 1991). Rosenthal often emphasizes that the BESD is a standard format for display of the Pearson correlation, and therefore, the propriety of reporting a BESD is conditional upon the propriety of reporting a Pearson correlation. Rosenthal notes that

> When used appropriately, the BESD has been used to excellent advantage by methodologically sophisticated behavioral researchers and by experienced mathematical statisticians . . . but we [Donald Rubin & Robert Rosenthal] are certainly agreed that the BESD is not the only way to tell how well we are doing in behavioral research. (p. 1087)

APPROPRIATE USE OF THE BESD

We believe that the BESD is a useful tool for reporting to evaluation audiences in two cases. The first case is when one wants to answer the questions that the BESD is meant to answer – "What would the correlationally equivalent effect of the treatment be if 50% of the participants had the occurrence and 50% did not and 50% received treatment and 50% did not?" This use of the BESD would be valuable when there is a need to illustrate that a value of *r* or r^2 that otherwise appears negligible may have practical significance, as was the case in the Vietnam and alcohol use example in Table 1. However, one must realize that the difference in percentages between a raw data two-by-two table and its corresponding *r*-based BESD is greatest when the symmetry of the cells in the raw data two-by-two table is lowest. In short, we agree with Rosenthal (1991) that there is no easy answer to which type of effect size to report under each of the various degrees of asymmetry.

The second case in which we believe the BESD is a useful tool is when there are not sufficient data to construct a raw-data two-by-two table. This often occurs with the results reported in meta-analyses. The authors of meta-analyses, and others who report on the results of a meta-analyses, often give effect sizes but do not, or cannot, report the aggregate, actual numbers of participants who improved or did not in each condition. Since effect sizes without raw data (e.g., when the actual proportion of the treatment group that improved is not known) abound in meta-analytic reports of research, the BESD can be put to good use to 'reframe' those effect sizes in an intuitive binomial display. To use an anecdotal example, one of the authors attempted to report the results of a meta-analysis on after-school research to a group of evaluation stakeholders who were considering planning and evaluating their own after-school program. In order to make their decision whether to implement the program, they wanted to know what kind of results, in terms of academic achievement, they could expect. There were no raw data reported in the meta-analysis; only that *d* was 0.13 in the after-school direction. After repeated attempts and diagrams to explain the interpretation of a *d* of 0.13, or an *r* of .065, the audience still was perplexed. However, after presenting the BESD that corresponds with a *d* or 0.13 (i.e., 53.5% of students improved in the after-school condition), the evaluation audience seemed to grasp how large an effect a *d* of .13 actually is. They commented that

the BESD should have just been shown in the first place.

CONCLUSION

Personal experience has shown us that statistically-laden effect sizes like *r* or *d* can be daunting for many evaluation audiences. However, most audiences seem to have little difficulty understanding percentages presented in two-by-two tables. In the absence of raw data, (e.g., an effect size from a meta-analysis) the BESD can be a useful substitute for showing the hypothetical magnitude of an effect if the assumptions of the BESD are addressed.

We agree with Rosenthal that "there is no right answer to which (indicator of effect size) is best or most useful under all conditions" (1991, p. 1086). Given the importance and difficulty of presenting statistically complicated results to lay evaluation audiences, it is useful for evaluators to report statistics, depending on the case, in many ways (e.g., difference in percentages, relative risk, odds ratios, proportions, standardized mean difference effect sizes, and correlational effect sizes). The BESD, when carefully used, is one of many ways that evaluators can put statistics to use to increase the utility value of program evaluation.

REFERENCES

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?". *American Psychologist, 46*(10), 1083.

May, H. (2004). Making statistics more meaningful for policy and research and program evaluation. *American Journal of Program Evaluation, 25,* 525-540.

McGraw, K. O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?". *American Psychologist, 46*(10), 1084-1086.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775-777.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist, 46*(10), 1086-1087.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3,* 377-386.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude and experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge: Cambridge University Press.

Strahan, R. F. (1991). Remarks on the binary effect size display. *American Psychologist, 46*(10), 1083-1084.

Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics, 22*(1), 109-117.

Worthen, R. B., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York: Longman.

Note

A previous version of this paper was delivered at the Annual Meeting of the American Evaluation Association; Reno, Nevada, November 5th through 8th, 2003.

Citation

Randolph, Justus J. & R. Shawn Edmondson (2005). Using the Binomial Effect Size Display (BESD) to Present the Magnitude of Effect Sizes to the Evaluation Audience. *Practical Assessment Research & Evaluation*, 10(14). Available online: http://pareonline.net/getvn.asp?v=10&n=14

Authors

Justus J. Randolph, Department of Computer Science, University of Joensuu, Finland; R. Shawn Edmondson, Department of Psychology, Utah State University.

Correspondence concerning this article should be addressed to Justus J. Randolph, Department of Computer Science, University of Joensuu, PO BOX 111, FIN-80101, Finland. E-mail: justus.randolph@cs.joensuu.fi