Calculating and Synthesizing Effect Sizes

Article *in* Contemporary issues in communication science and disorders: CICSD · January 2006

CITATIONS
56

READS
888

2 authors:

Herbert M. Turner Phd
University of Pennsylvania
17 PUBLICATIONS **781** CITATIONS
SEE PROFILE

Robert M Bernard
Concordia University Montreal
105 PUBLICATIONS **3,916** CITATIONS
SEE PROFILE

CICSD

Calculating and Synthesizing Effect Sizes

Herbert M. Turner, III
University of Pennsylvania, Philadelphia

Robert M. Bernard
Concordia University, Montreal, Quebec, Canada

It is assumed that at this stage in the systematic review on behavioral-based stuttering interventions, the authors have formulated the problem; implemented a thorough and comprehensive search; selected studies according to a set of inclusion criteria that were developed a priori; and coded these studies for subject characteristics, outcome characteristics, and even the characteristics of the studies themselves. Table 1 shows the results from coding outcomes for the James (1976) study that was included in the review. This study randomly assigned people who stutter (PWS) to a time-out (TO) intervention group and to a control group in which PWS did not receive the TO or any other intervention. This study design is called a randomized controlled trial (RCT). When implemented with a sufficient number of participants and with fidelity, an RCT produces unbiased estimates of an intervention's effect (Campbell & Stanley, 1963; Boruch, 1997; Mosteller & Boruch, 2002; Shadish, Cook, & Campbell, 2002). To assess the impact of TO, two outcomes were measured: percentage of stuttered words (%STW) and syllables per minutes (SPM). Table 1 shows the mean, standard deviation, and sample size for each group and outcome measure.

The data in Table 1 convey an intriguing picture of the relationship between the TO intervention and the outcomes. For the %STW outcome, the TO group scored, on average, more than two percentage points higher than the control group. Conversely, for the SPM outcome, the TO group spoke 22 more SPM, on average, than the control group. On the basis of these results, what can a clinician and practitioner conclude about the effect of TO on speech patterns for PWS?

To answer this question, the next logical step in the review process is to compute an effect size for each outcome in the James (1976) study and other studies that were included in the review. An effect size, as the name

ABSTRACT: The effect size is a standardized, scale-free measure of the relative size of the effect of an intervention, and it has important and practical implications for clinicians in the speech and hearing field who are interested in estimating the effects of interventions. This article develops a conceptual interpretation of the effect size, makes explicit assumptions for its proper use in estimating the size of the effect of behavioral-based stuttering interventions, and explains how to compute the most commonly used effect sizes and their confidence intervals. The focus is on effect sizes for experimental studies on behavioral-based stuttering interventions that produce outcomes measured on a continuous scale. Also explained is how to synthesize these effect sizes across multiple studies to arrive at an average effect size and its confidence interval through a process called meta-analysis. Key assumptions that underlie the use of meta-analysis are explored, as are techniques for assessing whether the average effect size is representative of the multiple effect sizes from which it was derived. The article concludes with a summary of main points and enumerates additional resources for speech and hearing clinicians and practitioners to access and learn more about practical applications of effect sizes and their synthesis through meta-analysis.

KEY WORDS: systematic review, meta-analysis, effect size

1092-5171/06/3301-0042

Table 1. Effect size data for a study included in the stuttering interventions review.

Study	Comparison	Outcome	Time-out group			Control group		
			$\overline{X}_I$	$\sqrt{S_I^2}$	n_I	$\overline{X}_C$	$\sqrt{S_C^2}$	n_C
James, 1976	Time-out vs. Control	%stuttered words (%STW)	7.40	6.66	9	11.95	6.58	9
James, 1976	Time-out vs. Control	Syllables per minute (SPM)	117.00	50.20	9	94.80	15.50	9

Note. For the time-out group, $\overline{X}_I$ is the mean, $\sqrt{S_I^2}$ is the standard deviation, and n_I is the sample size. For the control group, $\overline{X}_C$ is the mean, $\sqrt{S_C^2}$ is the standard deviation, and n_C is the sample size.

implies, enables the clinician and practitioner to estimate the effect of the behavioral-based stuttering intervention on the speech outcome. Eventually, clinicians and practitioners will want to estimate the average effect size for all behavioral-based stuttering interventions across all studies included in the review. The estimation process is called meta-analysis (Glass, McGaw, & Smith, 1981). In this article, the effect size is first explored from a conceptual and applied perspective rather than from a statistically theoretical one. However, basic formulas from introductory statistics and some technical language are necessary to explain the underlying logic of effect size estimation and to maintain the statistical integrity of this estimation process.

To this end, we begin with a conceptual interpretation of the effect size and basic but important assumptions that underlie this interpretation. Making plain these assumptions is important because violations of them can lead to misinterpretation of the effect size (Kline, 2004; Shavelson, 1996). Second, we explain how to compute the most commonly used effect sizes for study outcomes measured on a continuous scale. There are, of course, commonly used effect sizes for study outcomes measured on nominal scales. However, we restrict our explanation of effect sizes to study outcomes measured on a continuous scale in order to align the explanation with the type of outcomes reported in the studies included in the systematic review of behaviorally based stuttering interventions. Readers with an interest in effect sizes for outcomes measured on categorical scales are referred to Kline (2004) and Lipsey and Wilson (2001) as excellent sources. Third, we explain the process of meta-analysis and, specifically, how to aggregate effect sizes across studies to arrive at an average effect size. Meta-analysis is a critical stage of the systematic review process because the average effect size that is produced by a meta-analysis is interpreted as the average size of the intervention effect. We also explain how to compute the margin of error (or standard error) for this average intervention effect. We conclude by examining techniques for assessing whether a single average effect size is representative of the multiple effect sizes from which it was derived. We also summarize key ideas presented in this article and refer the reader to additional resources for understanding, estimating, and synthesizing effect sizes.

A CONCEPTUAL OVERVIEW OF AN EFFECT SIZE

The use of the effect size has grown significantly during the past three decades (Hunt, 1997; Hunter & Schmidt, 2004). Even before this growth, social scientists who were interested in estimating the effect of interventions viewed the effect size as a simple-to-calculate and useful representation of an intervention's effect (Cooper & Hedges, 1994; Light & Pillemer, 1984; Wolf, 1986). Today, the use of the effect size is generally accepted among social scientists to the point that its use is endorsed by the American Psychological Association (APA) (Kline, 2004).

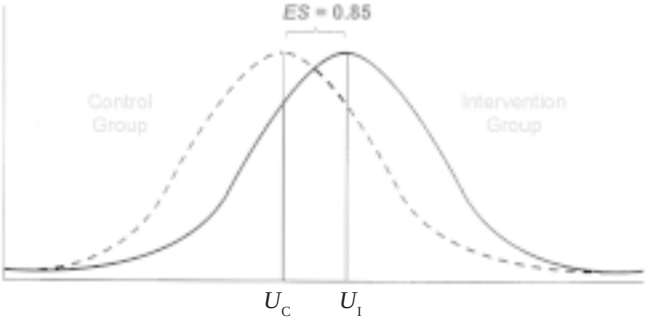
What is an effect size? Cohen (1988) defined an effect size as the degree to which the phenomenon is present in the population. Kline (2004) defined the effect size as the magnitude of the impact of the intervention on the outcome. Said differently, an effect size is an index for describing the magnitude of an intervention's effect. The effect size is recognized by researchers in a variety of disciplines as a simple and straightforward way to quantify the effects of an intervention relative to some benchmark comparison (Coe, 2002).

Figure 1 illustrates how the proper use and interpretation of the effect size is predicated on two basic but important assumptions. One assumption is that the frequency by which the outcome occurs is measured on a continuous scale and is normally distributed. An effect size estimate based on this assumption is technically called a parametric effect size because its estimation is based on the assumption of an outcome measure that is normally distributed (Glass et al., 1981; Shavelson, 1996).

Another assumption, also illustrated in Figure 1, is that the variance around the mean for the intervention group (U_I) and for the control group (U_C) is the same for each group. Therefore, when computing an effect size, homogeneity of group variances is assumed (Keppel & Wickens, 2004; Shavelson, 1996). Hereafter, when computing effect sizes, it is assumed that these two assumptions have been met. Approaches for addressing violations of these assumptions will be discussed later in the section on potential complications in computing effect sizes.

Conceptually, then, an effect size is an index that is used to describe the magnitude of an intervention's relative

Figure 1. Large effect size (ES) for intervention and control groups.



effect. Coe (2002) asserted, and we concur, that an effect size is readily understood and applicable to many measured outcomes in the behavioral sciences. This applicability extends with like force to the scientific investigations of speech, hearing, and communicative disorders. Equally important is that the effect size places the emphasis where it belongs—on the size of the intervention's effect rather than on its statistical significance. The problem with relying on statistical significance alone when interpreting the effect size is that statistical significance intermingles the effect size and sample size. How to disentangle effect size and sample size using confidence intervals will be discussed later on in the section on computing the margin of error for an effect size (Chalmers, Hedges, & Cooper, 2002; Kline, 2004; Light & Pillemer, 1984). How important is the use of effect sizes in the psychological and behavioral sciences? So much so that the American Statistical Association and the APA have recommended it as an important metric for conveying the quantitative relationship between an intervention and target outcomes (APA, 2001; American Statistical Association, 1992; Kline, 2004).

THE LOGIC OF EFFECT SIZE ESTIMATION

Types of Effect Sizes

There are many different types of effect sizes to choose from and books that describe them (Lipsey & Wilson, 2001; Rosenthal, 1984; Wolf, 1986; Hunter & Schmidt, 1982). There are also many different ways to calculate them (Hedges, Shymansky, & Woodworth, 1989; Light & Pillemer, 1984; Rosenthal, 1984). One way to organize the different types is to think of effect sizes as consisting of two broad families of indices. One family consists of the standardized mean difference, or *d* index, family. The *d* index, or standardized mean difference, is used to estimate the relative effect of an intervention for measure outcomes on a continuous scale. The other family consists of measures of association, or *r* family (Kline, 2004; Rosenthal, 1984). The *r* family is used to estimate the correlation between the intervention and the outcome. In practice, for studies that measure outcomes on a continuous scale, researchers often use the standardized mean difference and correlation coefficient because their standard error formulas and other useful statistical properties have been assessed and are well established in statistical theory (Kline, 2004; Lipsey & Wilson, 2001).

The focus of this section, and this article, is on computing the standardized mean difference within the *d* index family. This is because all of the studies included in the systematic review on behavioral-based stuttering interventions use an RCT design, lend themselves to a causal interpretation, and measure outcomes on a continuous scale (Boruch, 1997; Campbell & Stanley, 1963; Mosteller & Boruch, 2002; Shadish et al., 2002). Readers who are interested in computing effects in the *r* family are referred to Hunter and Schmidt (2004), Kline (2004), and Lipsey and Wilson (2001).

Estimating the Standardized Mean Difference

Recall that an effect size is defined as the degree to which the phenomenon is present in the population (Cohen, 1988). Arithmetically, the effect size (δ_i) is the ratio of the difference of the means of intervention and control groups ($\mu_I - \mu_C$) on an outcome divided by the standard deviation (σ_*). Thus, the formula for the standardized mean difference in the population is as follows:

$$\delta = \frac{\mu_I - \mu_C}{\sigma_*} \quad (1)$$

As Formula 1 shows, the effect size (δ) is expressed in standard deviation units because the divisor is the standard deviation (Cooper, 1998; Kline, 2004). However, it is only in rare circumstances that researchers have sufficient resources to study a population. Rather, researchers usually have sufficient resources to conduct an intervention study on a *sample* drawn from a population. Thus, in practice, the standardized mean difference in the population, δ, is estimated from sample data with the *d* index. The numerator of the *d* index ($\overline{X}_I - \overline{X}_C$) estimates the mean difference between the intervention and control groups in the population. The denominator (s^*) estimates the standard deviation in the population. Using sample data, the following formula is used to compute the *d* index (and estimator of δ in the population):

$$d = \frac{\overline{X}_I - \overline{X}_C}{s^*} \quad (2)$$

Hereafter, the standardized mean difference is referred to more generally as the *d* index.

Formula 2 plainly shows that the *d index* is the ratio of the difference between the *sample* mean for the intervention group and the *sample* mean for the control (or comparison) group, divided by the sample standard deviation. The standard deviation converts the mean difference in the numerator ($\overline{X}_I - \overline{X}_C$) into standard deviation units. Furthermore, the *d* index is an estimate of not one but two population parameters: the mean difference and the standard

deviation (Hedges & Olkin, 1985; Shavelson, 1996; Wolf, 1986). Effect sizes most commonly used in the behavioral sciences are Glass's Δ, Cohen's *d*, and Hedges's *g*. The main difference between Glass's Δ and Cohen's *d* is the standard deviation used in the denominator. Glass's Δ uses the standard deviation of the control group to standardize the mean difference between the intervention and control groups, as shown in Formula 3:

$$\text{Glass's } \Delta = \frac{\overline{X}_I - \overline{X}_C}{s_C} \tag{3}$$

Cohen's *d* uses the weighted average of the standard deviations of both groups to standardize the mean difference between the intervention and control groups, as shown in Formula 4.

$$\text{Cohen's } d = \frac{\overline{X}_I - \overline{X}_C}{\sqrt{\frac{(n_I - 1)s_I^2 + (n_C - 1)s_C^2}{n_I + n_C - 2}}} \tag{4}$$

The main difference between Cohen's *d* and Hedges's *g* is that the latter is multiplied by a correction factor for small samples.

$$\text{Hedges's } g = \text{Cohen's } d \bullet \left(1 - \frac{3}{4(n_I + n_C) - 9}\right) \tag{5}$$

Formula 5 shows that Hedges's *g* is actually Cohen's *d*—which uses the pooled standard deviations of both groups in the denominator—with a correction factor for use with small sample sizes. For example, to compute Cohen's *d*, using the data from the James (1976) study (see Table 1):

$$\text{Glass's } \Delta = \frac{\overline{X}_I - \overline{X}_C}{\sqrt{s_C^2}} = \frac{117.00 - 94.80}{15.50} = \frac{22.20}{15.50} = 1.43\ SD$$

This *d* index in the form of Glass's Δ says that the TO intervention caused a 1.43 *SD* increase in SPM. In other words, PWS in the intervention group spoke more SPM than did those in the control group. Notice that the result is reported in standard deviation units because the mean difference between the two groups was divided by the control group standard deviation. However, the standard deviation for the control group is three times smaller than that for the TO group. When the standard deviations for both groups are combined through Cohen's *d*, the effect size becomes:

$$d = \frac{\overline{X}_I - \overline{X}_C}{\sqrt{\frac{(n_I - 1)s_I^2 + (s_C - 1)s_C^2}{n_I + n_C - 2}}} =$$

$$\frac{117.00 - 94.80}{\sqrt{\frac{(9 - 1)(50.20)^2 + (9-1)(15.50)^2}{9 + 9 - 2}}} = \frac{22.20}{37.15} = 0.60\ SD$$

This *d* index in the form of Cohen's *d* says that the TO intervention caused a 0.60 *SD* increase in SPM. Notice that Glass's Δ is more than twice the size of Cohen's *d* because of the smaller standard deviation used to standardized Glass's Δ. Notice also that causal language is used to describe the effect of the TO intervention. Attaching a causal interpretation to the effect size is not because of anything inherent in the effect size itself. Rather, the causal interpretation of the effect is possible because study participants were randomly assigned to the TO and control groups (Boruch, 1997; Campbell & Stanley, 1963; Friedman, Furberg, & Demets, 1998; Mosteller & Boruch, 2002).

In practice, Cohen's *d* is used more often than Glass's Δ. However, when the combined sample sizes for the intervention and control groups are less than or equal to 20 ($n \leq 20$), Cohen's *d* is an upwardly biased estimator of the effect size in the population (δ) (Hedges & Olkin, 1985; Hedges et al., 1989). In other words, Cohen's *d* tends to overestimate an intervention's effect in small samples. To correct for the upward bias in Cohen's *d*, use Hedges's *g* (by applying Formula 5) as follows:

$$\text{Hedges's } g = \text{Cohen's } d \bullet \left(1 - \frac{3}{4N - 9}\right) =$$

$$0.60 \bullet \left(1 - \frac{3}{4 \bullet 18 - 9}\right) = 0.60 \bullet .95 = 0.57\ SD$$

The effect size estimated using Hedges's *g* (0.57) is smaller than the effect size that was estimated using Cohen's *d* because Hedges's *g* corrects for the upward bias that arises in Cohen's *d* when estimated in small samples. In practice, Hedges's *g* can be used in both large and small samples to avoid having to switch between Hedges's *g* and Cohen's *d* when a systematic review includes some studies that have small samples and others that have large samples (Chalmers & Altman, 1995). It has been demonstrated that Hedges's *g* converges to Cohen's *d* in large samples where $n > 20$ (Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Kline, 2004; Lipsey & Wilson, 2001). Like Cohen's *d*, Hedges's *g* has the following properties:

- It indexes the difference between the mean of the intervention group and the mean of the control group.
- It can be positive or negative.
- It is interpreted as a *z* score in standard deviation units. However, it should be noted that individual effect sizes are not part of the *z* score distribution (i.e., they will not sum to zero).

Recommendations for Clinicians and Practitioners

An effect size, as the name implies, conveys an estimate of the relative effect of an intervention. Use of the effect size has grown significantly during the past three decades.

Clinicians and practitioners who are interested in what works in speech and communication disorders should consider the APA Task Force's endorsement of the effect size and incorporate its use into their clinical work and

practice. When the variances between the groups compared are substantially different (i.e., heterogeneous), Glass's Δ should be used because it uses the control group standard deviation only to standardize the mean difference between two groups. Standard tests of homogeneity of variance can be used to determine this. When the variances of the groups are similar (i.e., homogenous), then either Cohen's *d* or Hedges's *g* should be used. It has been demonstrated that in small samples, Hedges's *g* is a better estimate of an intervention's effect in the population; it is as good an estimate as Cohen's *d* in large sample sizes. Therefore, Hedges's *g* is recommended for use in systematic reviews that include studies with both large and small samples or for use in clinical practice to compute effect sizes for groups in which the combined sample size is less than or equal to 20. Whichever effect size is used, the key point to remember is that an effect size is an *estimate* of an intervention's relative effect in the population from which the sample was drawn. For clinicians and practitioners who are interested in such, the effect size is available for use in a variety of practical settings.

COMPUTING MARGINS OF ERROR FOR EFFECT SIZES

Beyond Statistical Significance Testing

When computing Hedges's *g* for group comparisons such as the TO versus control comparison reported in the James (1976) study, the effect size may show a difference between the groups on the outcome of interest. This difference, however, could result from drawing the particular sample (for the intervention and control groups) from a population. In other words, the difference between the two groups that was observed in the sample may *not* exist in the population. How can one know whether a difference that was observed in a sample exists also in the population? To reduce the likelihood of reaching the wrong conclusion about an effect size in a population based on a sample, researchers have historically relied on "statistical significance" testing (Hunt, 1997; Keppel & Wickens, 2004; Kline, 2004). Statistical significance testing is defined as the likelihood that the observed difference between two groups *cannot* be attributed to chance (Hunt, 1997; Shavelson, 1996).

By convention, researchers usually set $\alpha = 0.05$ when testing a hypothesis about an intervention effect because $\alpha = 0.05$ is the probability that the observed difference being due to chance alone is equal to or less than 5%. When an RCT shows that the probability of a group difference is less than .05, the result is said to be statistically significant. That is, the observed mean difference is not due to chance alone.

Social science researchers such as Hunter and Schmidt (2004) and Kline (2004) have thoroughly documented limitations of "significance tests" to rule out chance results. Other social science researchers like Thompson (1999) have argued persuasively for a complete ban of the use of significance tests. Although banning statistical significance tests could be viewed as extreme, there is growing consensus that a fundamental weakness of significance tests is that strict observance of the *p* value confounds effect size with sample size (Coe, 2000; Kline, 2004; Shavelson, 1996). For example, it is possible to obtain a statistically significant result when the effect size is *large* in a *small* sample or when the effect size is *small* in a *large* sample. Furthermore, use of the statistical significance alone tells the clinician and practitioner nothing about the magnitude of the effect size, nor does it tell the interpreter about the effect size's precision as an estimator of the intervention's effect in the population (Coe, 2000; Kline, 2004; Rothstein, Sutton, & Borenstein, 2005; Shadish et al., 2002).

To overcome the limitations of the statistical significance test, the recommended approach is to report the *d* index with an estimate of its likely margin of error (Cooper, 1998; Cooper & Hedges, 1994; Lipsey & Wilson, 2001). A *d* index that is calculated from a large sample is likely to be a more accurate estimate of the true intervention effect, δ, in the population than is a *d* index that is calculated from a small sample. To quantify this margin of error, researchers calculate something called a confidence interval, which, ironically, contains the same information as a significance test and more (Hunter & Schmidt, 2004; Kline, 2004; Thompson, 1999). In fact, calculating a 95% confidence interval for an observed *d* index is equivalent to calculating a 5% significance level in which $p \leq 5$ means that the *d* index is statistically significant (Shavelson, 1996). Why this is so and how to compute the margin of error (or standard error) and confidence interval for the *d* index is explored in the next section. For now, it is sufficient to say that computing the margin of error (or standard error) for the *d* index is equivalent to repeatedly drawing samples of the same size, computing the *d* index, and recording the range of results.

Although computing an effect size, its margin of error (or standard error), and its 95% confidence interval would seem tedious to do for multiple studies, there are a number of software packages, both free and commercial, that are available to reduce this tediousness.

Recommendations for Clinicians and Practitioners

The APA Task Force on Statistical Significance recommends that effect sizes, as well as other estimates of intervention effects, be reported with their confidence intervals. Speech and hearing clinicians and practitioners should adhere to this recommendation when assessing the effects of an intervention because in doing so it recognizes that the effect size is an estimate from sample data, the estimate contains sampling error, and a confidence interval distinguishes between an effect size and sample size (Clarke, 2002). Although computation of confidence intervals may appear tedious and complicated to compute for clinicians and practitioners in the speech and hearing field, there are a number of excellent computer software packages to aid in the process. Standard statistical software packages such as STATA (www.stata.com), SAS (www.sas.com), and SPSS (www.spss.com) can be used to

easily compute effect sizes and their confidence interval. However, these software programs (STATA and SAS especially) are geared toward persons with a computer programming orientation. For clinicians and practitioners with a different orientation, there are user-friendly software packages such as CMA 2.0 (www.cma.com), RevMan 5.0 (www.cochrane.com), and Metawin (www.metawin.com) that offer as many options as standard statistical software packages and are easier to use (Borenstein, 2005; Borenstein & Rothstein, 1999). See Appendix A for a list of web sites that are rich resources for clinicians and practitioners to learn more about computing effect sizes, margins of error, and confidence intervals.

ESTIMATING THE EFFECT OF STUTTERING INTERVENTIONS

Estimating Intervention Effects Using Ryan (1995)

In this section, effect size concepts and formulas that were presented earlier are applied to outcomes that were coded in the stuttering review. One concern the reader may have at this point is the tediousness of computing Hedges's *g* and its confidence interval for all of the comparisons in all studies in the stuttering review. As mentioned earlier, there are a number of free and commercial software packages that can be used to implement these computations. These packages are listed in Appendix B.

Table 2 depicts outcome data for the Ryan (1995) study, which is one of the 12 RCTs included in the stuttering review. The first three columns specify the study author, group comparisons, and outcomes. The means ($\overline{X}$), standard deviations (*s*), and sample sizes (*n*) that are needed to compute effect sizes are presented in columns 4 through 9. The remaining columns contain the *d* index (g_H) and its standard error (SE_g). Table 2 shows that the Ryan RCT compared a group of study participants who received a gradual increase in length and complexity of utterance (GILCU) intervention to a group of study participants who received a delayed auditory feedback (DAF) intervention.

The Ryan (1995) study differs from the James (1976) study in that the former study was designed to test whether GILCU was more effective than DAF (i.e., the DAF group was a comparison rather than a control group) whereas the latter study was designed to test whether the TO intervention had an effect (i.e., the TO intervention group was compared to a control group who did not receive any intervention). Thus, for the Ryan study, the GILCU group was designated as the intervention group and the DAF group was designated as the control group because, in this example, the interest is in how effective the GILCU intervention was relative to the DAF intervention. There are multiple lines of data for the same comparison in this study because the same groups were compared on multiple outcomes, which were as follows:

- SW/M – stuttering words per minute
- WS/M – words spoken per minute

Notice that Table 2 shows that the combined sample sizes for the Ryan study are small ($n \leq 20$). Therefore, Hedges's *g* was used to compute the effect sizes. Applying Formula 5 to calculate Hedges's *g* for the SW/M outcome results in:

$$g_{\text{Hedges}} = \frac{0.7 - 0.5}{\sqrt{\frac{(5-1)*(0.5)^2 + (6-1)*(0.7)^2}{5+6-2}}} \cdot \left(1 - \frac{3}{4(5+6)-9}\right)$$

$$= 0.30$$

This effect size says that the GILCU intervention increased the number of spoken words per minute by 0.30 *SD*. The use of the causal language is approximate because the Ryan study was, as stated earlier, an RCT (Boruch, 1997; Campbell & Stanley, 1963). What would have been the result if Cohen's *d*, rather than Hedges's *g*, had been used? As stated earlier, Cohen's *d* is really Hedges's *g* but without the small sample correction. To compute Cohen's *d*, we have:

$$d_{\text{cohen}} = \frac{0.7 - 0.5}{\sqrt{\frac{(5-1)*(0.5)^2 + (6-1)*(0.7)^2}{5+6-2}}} = 0.32$$

As expected, Cohen's *d* is slightly larger than Hedges's *g* because the former is upwardly biased in small samples.

Table 2. Outcome data coded for a subset of studies from the stuttering review.

			GILCU group			*DAF group*			*Effect size estimate*	
Study	*Comparisons*	*Outcomes*	$\overline{X}_I$	$\sqrt{S_I^2}$	n_I	$\overline{X}_C$	$\sqrt{S_C^2}$	n_C	g_H	SE_g
Ryan, 1995	DAF vs. GILCU	SW/M	0.7	0.7	6	0.5	0.5	5	0.30	0.56
Ryan, 1995	DAF vs. GILCU	WS/M	137.5	35.3	6	133.4	17.4	5	0.13	0.55

Note. DAF = delayed auditory feedback; GILCU = gradual increase in length and complexity; $\overline{X}_I$ = GILCU group mean; S_I = GILCU group standard deviation; n_I = GILCU group sample size; $\overline{X}_C$ = DAF group mean; S_C = DAF group standard deviation; n_C = DAF group sample size; SW/M = stuttering words per minute; WS/M = words spoken per minute.

Calculating The Margin of Error in Effect Sizes

As stated earlier, values of Cohen's *d* and Hedges's *g* are estimates of an unknown effect size in the population (δ). For instance, the Hedges's *g* just computed will, theoretically, change with repeated samples of the same size drawn from the population if the Ryan study could be replicated. The standard error produces the range of values that would result from this process of repeated sampling. For example, to compute the standard for the Hedges's *g* of 0.30, the following formula (Formula 6) is used:

$$(6)\ SE_g = SE_d \bullet j = \sqrt{\frac{1}{n_I} + \frac{1}{n_C} + \frac{d_{cohen}^{\ 2}}{2(n_I + n_C)}} \bullet \left(1 - \frac{3}{4(n_I + n_C) - 9}\right) =$$

$$\sqrt{\frac{1}{5} + \frac{1}{6} + \frac{-0.32^2}{2(6 + 6)}} \bullet \left(1 - \frac{3}{4(6 + 6) - 9}\right) =$$

$$0.61 \bullet 0.91 = 0.56$$

The next step is to construct the lower and upper limits of approximate 95% confidence interval by setting $\alpha = .05$ and $z = 1.96$ such that:

$$\text{Lower limit} = \text{Hedges } g_{(U)} = \text{Hedges } g_{(U)} - Z_{(1 - .05)}$$

$$SE_{\text{Hedges's g}} = 0.32 - 1.96 \bullet 0.56 = -0.78$$

$$\text{Upper limit} = \text{Hedges } g_{(U)} = \text{Hedges } g_{(U)} - Z_{(1 - .05)}$$

$$SE_{\text{Hedges's g}} = 0.32 + 1.96 \bullet 0.56 = 1.42$$

This result says that our sample estimate of the effect of GILCU on study participants is 0.30 *SD*, but the effect of GILCU on study participants in the population (from which the sample was drawn) has a range (95% CI = –0.78, 1.42). This means that if repeated samples of the same size could be drawn for the Ryan study, these samples would contain observed values of Hedges's *g* that range from a low of –0.78 *SD* to a high of 1.42 *SD*. In practice, the primary concern is whether the confidence interval includes an effect size of 0.00 *SD* because even though a positive (or negative) effect size was observed in the sample, in the population, this effect size could be zero. In other words, if the confidence interval includes zero, the observed Hedges's *g* of 0.30 is not statistically significant because the observed value of 0.30 could be due to chance, meaning that the effect size is really 0.00 *SD*. In other words, the intervention has no effect.

One complication that arises quite frequently in practice is how to compute an effect size when study results reported are in formats other than means, standard deviations, and sample sizes. For example, in some studies, authors will present results from an analysis of variance (ANOVA) or an analysis of covariance (ANCOVA) using *F* ratios or sums of squares, or both. Furthermore, it is almost inevitable that a study included in a review will have missing values on the data needed to compute an effect size (e.g., the means, standard deviations, or sample sizes). The next section of this article addresses how to deal with these complications.

Potential Complications in Estimating the *d index*

Nonnormal distributions and unreliable outcome measures. Two important complications to consider when computing an effect size are nonnormal distributions and measurement reliability of the outcomes of interest. Recall that one of the assumptions for using the *d* index is that the distributions for two groups are normal (see Figure 1). When this is not the case, a nonparametric effect size must be used (Kline, 2004). However, use of the nonparametric effect size is beyond the scope of this article; the reader should see Kline (2004) and Hunter and Schmidt (2004), which provide in-depth treatments of the topic. However, in large samples, the *d* index is fairly robust to departures from normality in distributions (Kline, 2004; Shavelson, 1996). When the outcome is measured unreliably, the *d* index tends to be smaller than it otherwise would have been if the outcome had been measured reliably (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002).

Which* d *index to use? The formulas for computing the *d* index are suitable for most situations. There are some complications that often arise during the review. The first is the choice of which standard deviation to use in the denominator of the *d* index. Table 3 illustrates how this choice can impact the effect size.

Table 3 shows that the sample sizes in the Ryan study are by definition small ($n \leq 20$). This suggests that, all other things being equal, Hedges's *g* is the most appropriate effect size to use because of the small sample sizes. However, it is also important to consider how the standard deviations for each group compare before deciding on which effect size to use. For the SW/M outcome (Table 3, row 1), the sample sizes and standard deviations for the two groups are roughly equal, although it appears that the GILCU intervention also increased the average variability in the GILCU group. Still, there is not much difference between the three types of effect sizes, and because the sample sizes are small, the use of Hedges's *g* (g_H) is most appropriate. Similar decision logic can be applied to the data on the second outcome in the Ryan study, which is WS/M. For the third outcome in the Ryan study, the intervention has increased the variability in SW/M such that the variance of the intervention group is almost five times that of the control group. In this case, use of Glass's Δ would produce a much larger effect size because its smaller denominator increases the size of the intervention effect.

Different formats for effect size data. It is not unusual for a review to include studies that report outcomes on a continuous scale but in nonconventional formats. Nonconventional formats are defined as outcome data that are reported in formats other than means, standard

Table 3. Effect sizes for the DAF vs. GILCU comparison on three outcomes in one study.

		Intervention group			*Control group*			*Effect sizes*		
Study	*Outcome*	$\overline{X}_I$	S_I^2	n_I	$\overline{X}_C$	S_C^2	n_I	*Glass's* Δ	*Cohen's d*	*Hedges's g*
Ryan, 1995	SW/M	0.70	0.70	6	0.50	0.50	5	0.29	0.32	0.30
Ryan, 1995	WS/M	137.50	35.30	6	133.40	17.40	5	0.12	0.14	0.13
Ryan, 1995	SW/M FU	0.60	0.30	6	1.10	1.70	5	1.67	0.43	0.40

deviations, and sample sizes. Examples of nonconventional data formats, from which effect sizes can be derived for posttest data on unmatched groups, is as follows:

- Means, sample size, and *t* value (or two-group *F* ratio)
- Sample size and *t* value (or two-group *F* ratio)
- Means, sample size, and *p* value
- Sample size and *t* value (or two-group *F* ratio)

For these data, one can apply the derivation formulas to compute Hedges's *g* or Cohen's *d*. Although these derivations can be done manually or by using a calculator, the derivation formulas are usually implemented using computer software. To learn more about these formulas and their use, refer to Borenstein (2005), Hunter and Schmidt (2004), Kline (2004), and Lipsey and Wilson (2001).

Missing values on outcome data for effect sizes. It is also not unusual for an analyst to encounter studies with missing values on the outcome data that were used for effect size computations. Referring to Table 3 again, if Ryan had only reported means and had not reported the standard deviations for the outcomes, it would be impossible to compute the effect size. In another study included in the stuttering review but not shown in Table 3, the author reported means and samples sizes but not the standard deviations, and another study author reported *F* ratios and *p* values rather than the means, standard deviations, and sample sizes. Under these circumstances, and others like them, there are four options for computing effect sizes:

Contact study authors. Contact the study authors and request the missing data. This option is the best among the four proposed, but successful implementation depends on, among other things, how recently the study was reported. Authors of older studies are more difficult to locate and contact than are authors of more recent studies. Furthermore, even if the authors can be contacted, there is no guarantee as to whether they have the original data or, if they have the data, whether they are willing to release it.

Set the effect size to zero. When study authors cannot be contacted or are unwilling to release the requested data, then the value of the effect size, for which there are missing values that prohibit its calculation, can be set to zero. Setting the missing effect size equal to zero is viewed as establishing a conservative estimate of the missing effect size value. Researchers are often uncomfortable with such a conservative estimate because it results in an effect size that theoretically could be smaller or larger than it would have been if the effect size could have been computed from the available data. Because it also reduces sample variance, this approach should be used sparingly.

Use multiple imputation. A more sophisticated (and complicated) approach is to use multiple imputation to impute the missing effect sizes using the nonmissing effect sizes in the review. Multiple imputation is the process of replacing missing effect sizes with the average of effect sizes drawn from a random distribution of imputed effect sizes derived from the observed effect sizes (Allison, 2001). Multiple imputation is recommended over mean imputation (where the mean value of the nonmissing effect sizes is substituted for the missing effect sizes) and other types of imputation because multiple imputation produces better estimates of the true effect size (Allison, 2001; Pigott, 2001). The use of multiple imputation is beyond the scope of this article, but the reader is referred to Allison and Pigott to learn more about the theory behind and implementation of multiple imputation.

Omit the study. When the first three options are unappealing or cannot be implemented, then the last option is to omit the study from meta-analysis. When this option is exercised, it reduces the number of studies in the meta-analysis by S – M, where S is the total number of studies included in the review and M is the number of studies excluded from the meta-analysis because of missing values on the outcome data used to compute effect sizes.

Recommendations for Estimating Effect Sizes in Practice

Use Hedges's* g *when feasible. When the frequency of an outcome for the intervention and control groups is normally distributed on an outcome that is measured on a continuous scale, and the variances of the two groups are similar, Hedges's *g* or Cohen's *d* should be used to estimate intervention effects. Furthermore, when the systematic review consists of large and small sample studies, Hedges's *g* should be used because it corrects for the small sample bias in Cohen's *d* and converges to Cohen's *d* in large samples. In situations where the variance of the intervention group is altered by the treatment to the point that there is a large disparity between this variance and that of the control group, then Glass's Δ should be used. However, use of Glass's Δ should be weighed against other factors such as whether the disparity in the variances between the two

groups is a valid consideration as part of the intervention effect. If so, then Hedge's *g* is a better index because it pools the variance of both intervention and control groups. In cases where the assumption of normality is violated, nonparametric effect sizes must be used (Kline, 2004).

Use means, standard deviations, and sample sizes when feasible. It is not unusual for study authors to report outcome data in formats other than means, standard deviations, and sample sizes. Fortunately, the formulas for deriving effect sizes from various data formats are well established and can be implemented using computer software. Still, when possible, it is better to compute effect sizes from means, standard deviations, and sample sizes even when it requires the extra effort of contacting study authors. When missing values cannot be obtained by contacting study authors, one recommended approach is to set the effect size for the group comparison equal to zero. A better but more complicated approach is to use multiple imputation to impute the missing effect size values using the nonmissing effect size values.

Report confidence intervals when feasible. The effect size should be reported with its confidence interval because it is an estimate of a difference in the population based on a sample (APA, 2001; American Statistical Association, 1992). The confidence interval displays the amount of error in this estimate. In practical terms, this confidence interval displays the range of plausible intervention effects (or effect sizes) in the population from which the intervention and comparison groups were selected.

SYNTHESIZING THE *d* INDEX ACROSS STUTTERING INTERVENTION STUDIES

Steps for Synthesizing the *d* Index

Having computed effect sizes for all comparisons reported in included studies, the next step in the systematic review process is to estimate the average effect size. The average effect size is estimated by statistically synthesizing the effect sizes across the studies. This process is called meta-analysis (Cooper & Hedges, 1994; Glass et al., 1982; Hunt, 1997). Meta-analysis can be defined as the statistical synthesis of the data from separate but comparable studies leading to a quantitative summary of pooled results (Last [2001] in Chalmers et al., 2002). The focus of this section is on how to conduct a bare-bones meta-analysis. A bare-bones meta-analysis is the first stage of a complete meta-analysis and is called bare bones because it controls only for variability in effect sizes due to the sampling of subjects within each study (i.e., sampling error) (Hunter & Schmidt, 2004). Advanced stages of a meta-analysis are designed to control for variability in effect sizes due to other factors like subgroup or study characteristics. These stages are beyond the scope of this article but are discussed extensively in Hunter and Schmidt, Kline (2004), and Lipsey and Wilson (2001). The main steps for a bare-bones meta-analysis are as follows:

1. Average effect sizes within studies (if necessary).
2. Average effect sizes across studies.
3. Assess observed variability in effect sizes pooled across studies.
4. Interpret overall average effect size.

Implementation of each step is discussed next.

Averaging Effect Sizes Within Studies

Table 4 shows the six included studies for the stuttering systematic review along with their comparisons, outcomes, group means, standard deviations, and sample sizes. Four of the six studies each have a combined sample size (intervention and control group) of less than 20 study participants. For this reason, Hedges's *g* was used to estimate the effect sizes (Table 4). Formulas 5 and 6 were used to compute the effect sizes and their standard errors, respectively. Table 4 also shows that some studies report multiple outcomes for a single comparison, such as Boudreau (1973) and Waterloo (1988), whereas other studies, such as Öst (1976) and Harris (2002), report a single outcome for a single comparison. Studies of the latter can be included as in the meta-analysis. Studies of the former cannot because the same participants who are in the intervention and control groups for one outcome are the same people who are in the intervention and control group for another outcome. Including the same comparison on multiple outcomes in the meta-analysis violates the assumption of statistical independence and renders the standard errors and confidence intervals inaccurate (Cooper, 1998; Cooper & Hedges, 1994; Hunter & Schmidt, 2004; Kline, 2004; Lipsey & Wilson, 2001).

To avoid violating the assumption of independence of effect sizes, reviewers will implement one of the following three approaches:

- Randomly select one effect size.
- Select one effect size based on theory or a logically defensible rationale.
- Compute a simple average of the effect sizes.

Reviewers in practice usually implement the third approach and compute an average of the effect sizes (and their standard errors) unless the outcome measures are so conceptually different that it is inappropriate. For example, the simple average of the four effect sizes computed for the Boudreau (1973) study is the sum of the four effect sizes divided by that number of effect sizes ($n = 4$), resulting in an average Hedges's *g* of 0.57. The disadvantage of this approach is that it becomes more difficult to code study features (e.g., setting, grade) later on because now the effect size represents an average of two conditions that may have different study characteristics.

A simple average of the four standard errors for the four effect sizes results in an average standard error of 0.55. The results from averaging the effect sizes and standard errors for the other studies for which multiple effect sizes were computed (e.g., James, 1976; Waterloo, 1988) are presented in Table 5. Note that studies with one outcome and hence one effect size were not averaged. Fortunately, meta-analysis software such as CMA 2.0 will compute the

Table 4. Effect size data for studies included in the stuttering interventions review.

			Intervention group			*Control group*			*Effect size estimate*	
Study	*Comparison*	*Outcome*	$\overline{X}_I$	$\sqrt{S_I^2}$	n	$\overline{X}_C$	$\sqrt{S_C^2}$	n	g_{Hedges}	SE_g
Boudreau, 1973	Desen vs. Ctrl	%SS	6.38	9.21	12	11.75	8.06	4	0.57	0.55
Boudreau, 1973	Desen vs. Ctrl	%SW RA	11.25	13.19	12	17.00	9.45	4	0.44	0.55
Boudreau, 1973	Desen vs. Ctrl	%SW SS A	9.13	7.26	12	16.00	8.87	4	0.85	0.57
Boudreau, 1973	Desen vs. Ctrl	%SWSS AC	11.63	6.91	12	28.75	16.21	4	1.67	0.62
Öst, 1976	Desen vs. Ctrl	%NFL SS	9.30	8.40	5	12.80	16.80	5	0.24	0.57
James, 1976	Time-out vs. Ctrl	%SS	7.40	6.66	9	11.95	6.58	9	0.65	0.46
James, 1976	Time-out vs. Ctrl	SPM	117.00	50.20	9	94.80	15.50	9	0.57	0.46
Waterloo, 1988	Reg. Breath vs. Ctrl	%SW Read	6.80	6.80	16	18.50	8.40	15	1.50	0.40
Waterloo, 1988	Reg. Breath vs. Ctrl	%SW SS	4.50	5.70	16	17.90	6.70	15	2.10	0.44
Waterloo, 1988	Reg. Breath vs. Ctrl	%SW Pho	5.80	5.80	16	20.10	8.00	15	2.00	0.43
Harris, 2002	Lidcombe vs. Ctrl	%SS	3.50	2.80	8	5.80	3.60	11	0.67	0.46
Jones, 2005	Lidcombe vs. Ctrl	%SS	1.50	1.40	27	3.90	3.50	20	0.94	0.31

Note. Desen = desensitization intervention group, Ctrl = control group, Reg. Breath = regulated breathing intervention group, %SS = percentage of stuttered syllables, %SW = percentage of stuttered words, %NFL = percentage of nonfluency, SPM = stuttered words per minute.

average for you (see Appendix B). With one effect size per study, the next step is to pool the effect sizes to arrive at the average effect size across the five studies.

Averaging Effect Sizes Across Studies

Although a simple average was appropriate for combining multiple effect sizes *within* studies, a weighted average is used to pool effect sizes *across* studies. An effect size from a study with a larger sample, like the Jones (2005) study, with 47 participants, provides a more precise effect size estimate (i.e., a lower standard error of 0.31 and, hence, a smaller confidence interval) than a study with a small sample, like the Öst (1976) study, with only 10 participants (and a higher standard error of 0.57 and, hence, a larger confidence interval). To weight each effect size, simply square the standard error to get the variance and then compute its inverse or reciprocal (i.e., place one over the squared standard error) as follows:

$$(7) \qquad IVW_{Hedges's\ g} = \frac{1}{(SE_{Hedges's\ g})_2} = \frac{1}{(0.574)^2} = \frac{1}{0.33} = 3.03$$

Table 6 shows the included studies with their effect sizes, the inverse variance weight for Hedges's g ($IVW_{Hedges's\ g}$), and the effect sizes weighted by the $IVW_{Hedges's\ g}$.

The data in Table 6 are all that is needed to compute the average effect size. To compute the average effect size, sum the five weighted effect sizes and divide by the sum of the IVW. These two quantities are reported in the last two columns of the last row in Table 6. More formally, the formula for the average effect size is:

$$(8) \qquad \text{Hedges's } g^* = \frac{\sum_{i=1}^{n} \text{Hedges's g(i)} \bullet \text{IVWHedges's g(i)}}{\sum_{i=1}^{n} \text{IVWHedges's g(i)}}$$

In Formula 8, "i" is the number of study effect sizes; for this meta-analysis, the numerator and denominator in Formula 8 are summed across five studies. Substituting the values from the last row under the last two columns from Table 6 yields:

$$\text{Hedges's } g^* = \frac{27.22}{31.83} = .86$$

This result says that on average, the stuttering interventions caused more than a three quarters standard deviation improvement in the speech pattern of PWS. Furthermore, this effect, though quite large, is an estimate from a sample of studies and therefore its standard error should be used to compute an upper and lower confidence interval. The standard error can be easily computed using the sum of the inverse variance weights as follows:

Table 5. Average effect sizes and standard errors for studies included in the stuttering interventions review.

Study	Comparison	# of Outcomes per study	Outcome	Intervention group n	Control group n	Effect size estimate Hedges's g	Effect size estimate SE_g
Boudreau, 1973	Desen vs. Ctrl	4	Averaged	12	4	0.88	0.57
Öst, 1976	Desen vs. Ctrl	1	%NFL SS	5	5	0.24	0.57
James, 1976	Time-out vs. Ctrl	2	Averaged	9	9	0.04	0.46
Waterloo, 1988	Reg. Breath vs. Ctrl	3	Averaged	16	15	1.87	0.42
Harris, 2002	Lidcombe vs. Ctrl	1	%SS	8	11	0.67	0.46
Jones, 2005	Lidcombe vs. Ctrl	1	%SS	27	20	0.94	0.31

$$(9) \qquad SE_g = \sqrt{\frac{1}{\sum_{i=1}^{n} \text{WHedges's g(i)}}} = \frac{1}{31.83} = 0.18$$

The 95% confidence interval for the average effect size of 0.86 with a standard error of 0.18 is:

$$(10) \quad \text{Hedges's g} \times \pm 1.96(SE_{\text{Hedges's g*}}) = 0.86 \pm 1.96(0.18) =$$

$$0.86 \pm 0.35 = [1.20, 0.51]$$

The conclusion is that stuttering interventions reported in the review resulted in an improvement in speech patterns of PWS in the *sample* by approximately 0.86 *SD*. However, when this improvement is generalized to the sample from which study participants were drawn, this improvement could be as high as 1.20 *SD* or as low as 0.51 *SD*. Despite this variability (or error) in the estimate, the conclusion from this bare-bones meta-analysis is that one can be 95% confident that the effect of stuttering interventions is positive.

Assessing Observed Variability in Effect Sizes Pooled Across Studies

The results of the bare-bones meta-analysis can also be displayed graphically through a forest plot. A forest plot allows reviewers to separate the forest (the average effect and its standard error) from the trees (the study effect sizes and their standard errors) (Rothstein et al., 2005). The forest plot also allows reviewers to assess how representative the average effect size is of the study effect sizes because each study effect size and its 95% confidence interval is displayed relative to the average effect size and its 95% confidence interval. Any of the meta-analysis software mentioned earlier can be used to produce these plots. Figure 2 presents a forest plot of the effect size data from the stuttering review just discussed (Table 6).

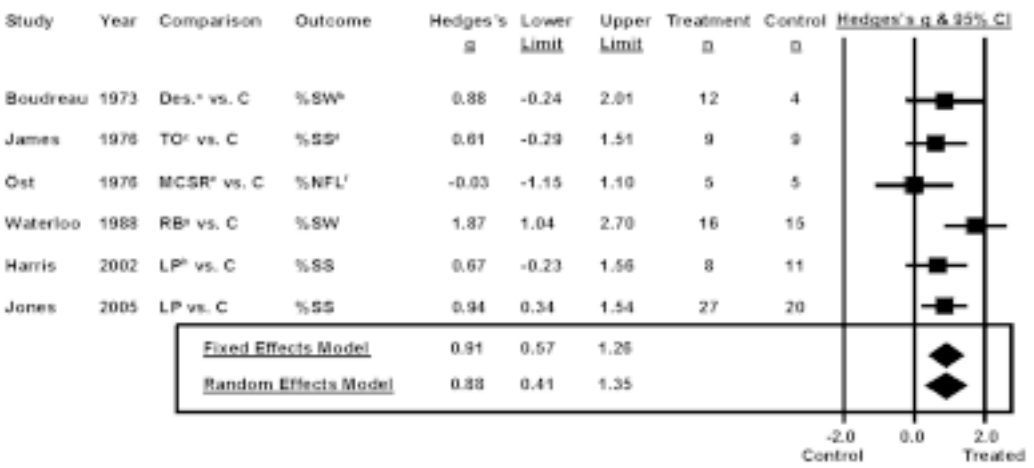
The forest plot illustrates how the study effect sizes vary relative to the average effect size of 0.86 *SD*. All of the studies have positive point estimates (squares in Figure 2) signifying that the behavioral intervention for stutterers is positive. However, these point estimates vary in their precision as the confidence intervals are wider for some estimates than for others (bars in Figure 2; more on this in the next section on interpreting effect sizes). The effect size for the Öst (1976) study is much smaller than the average effect size, but the effect size for the Waterloo (1988) study is much larger. The variability in these effect sizes begs the question: Is the average effect size representative of the study effect sizes from which it was derived?

This is an important clinical question because an estimated average treatment effect size that is not

Table 6. Weighted effect sizes for studies included in the stuttering interventions review.

Study	Comparison	Outcome	Intervention group n	Control group n	Weighted effect size estimates Hedges's g	$W_{\text{Hedges's g}}$	Hedges's g × $W_{\text{Hedges's g}}$
Boudreau, 1973	Desen vs. Ctrl	Averaged	12	4	0.88	3.04	0.57
Öst, 1976	Desen vs. Ctrl	%NFL SS	5	5	0.24	3.04	0.57
James, 1976	Time-out vs. Ctrl	Averaged	9	9	0.04	4.72	0.46
Waterloo, 1988	Reg. Breath vs. Ctrl	Averaged	16	15	1.87	5.56	0.42
Harris, 2002	Lidcombe vs. Ctrl	%SS	8	11	0.67	4.79	0.46
Jones, 2005	Lidcombe vs. Ctrl	%SS	27	20	0.94	10.69	0.31
The Summation (Σ)			77	64		31.83	27.22

Figure 2. Overall effect for behavioral stuttering intervention: Treatment vs. control.



Note. Heterogeneity statistics for a fixed model: Q = 8.49, *df* = 5, *p* = .13, I^2 = 41.07.

[a]Desensitization; [b]percentage of stuttered words; [c]Timeout from speaking; [d]percentage of stuttered syllables; [e]metronome-conditioned speech retraining; [f]percentage of nonfluency; [g]regulated breathing; [h]Lidcombe program.

representative of the study effect sizes from which it was derived could raise expectations for the behavioral intervention to have a certain level of impact that cannot be delivered because the study effect sizes varied so widely.

The Q and I^2 statistics are used to empirically assess the amount of variability in effect sizes beyond what is expected from sampling error (Higgins, Thompson, Deeks, & Altman, 2003). The Q statistic tests the null hypothesis that the effect sizes in the meta-analysis are estimating the same effect size in the population. Figure 2 shows that a Q statistic of 10.22 is not statistically significant (p = 0.69). This means that the variability in the effect sizes, as depicted in Figure 2, is what is expected given sampling error. However, the Q statistic is not reliable when it is computed based on a small number of studies, as is the case in this meta-analysis. This is one reason that the I^2 statistic should be interpreted along with the Q statistic. The I^2 statistic represents the amount of observed variation in effect sizes that is attributable to factors other than sampling error. The note under Figure 2 shows that I^2 = 51.08%, which means that approximately half of the observed variation is due to sampling error and the other half is due to other possibly unknown factors. When interpreting, the following guidelines for values of I^2 may be helpful (Higgins et al., 2003):

- 25% or less indicates small amounts of heterogeneity.
- 50% indicates moderate amounts of heterogeneity.
- 75% or more indicates large amount of heterogeneity.

Taken together, the results from this review say that there is 95% confidence that behavioral stuttering interventions have a positive effect. Furthermore, the effect is at least a half a standard deviation and could be larger, up to just more than 1.0 *SD*. However, this average effect was derived from a set of effect sizes that were moderately heterogeneous, meaning that the average effect could be larger, smaller, or stay the same if there are more studies that investigate these other factors. Still, given the size of the effect and its confidence intervals, there is enough evidence to be confident that this effect size would be positive even after such an investigation.

Recommendations for Synthesizing and Interpreting Effect Sizes in Practice

In practice, it is important to consider four elements in synthesizing effect sizes into a composite mean that can be interpreted as a general effect in the population.

- First, how within-group effects are treated is important and care should be taken in the approach that is used to combine several effect sizes within a given study. Effect sizes must be independent, that is, they must each contain different study participants. If there are multiple treatments and one control, for instance, the control participants cannot be used twice to form two effect sizes. One common way of solving this problem is to form a simple average of the two treatment effect sizes.
- Second, we need to average the effect sizes across the distribution of studies. This cannot be a simple average because of the effects of differential sample size. This special mean is called a weighted mean and takes into consideration the relative impact of large versus small samples.
- The third element is to find the standard error of the mean and use it to calculate the upper and lower bounds of the 95th percent confidence interval. This interval can be used to test the null hypothesis that effect size = 0. The confidence interval also describes the limits within which the true mean lies.

- The fourth aspect of importance is the variability that surrounds the average effect size. The Q statistic describes this variability and can be tested to determine if the effect sizes in the distribution are homogeneous (i.e., are likely to correctly describe the average effect in the population) or heterogeneous (i.e., are unlikely to correctly describe the average effect size in the population). If homogeneity of effect size cannot be achieved, strong statements about the effect of the independent variable on the dependent variable in the population are not warranted. In this case, the researcher may want to explore the effects of moderator variables in an attempt to achieve homogeneity of effect size. The I^2 statistic is useful as a supplement to Q in interpreting the variability surrounding the mean effect size.

SUMMARY

The effect size is a standardized, scale-free estimate of the relative size of the effect of an intervention in the population. Accurate interpretation of the effect size rests on the assumption that the intervention and control groups are normally distributed on an outcome variable and that these groups have the same standard deviations (i.e., homogeneity of variances). In research that uses an RCT and the outcomes measured on a continuous scale, the *d* index in general and Hedges's *g* is the effect size that is used most often. Use of an effect size with its confidence interval expresses the same information as a test of statistical significance but places the emphasis on the size and significance of the effect rather than on a sample size. For this reason, when feasible, effect sizes should always be reported with their confidence intervals. This applies to the reporting of effect sizes in primary studies as well as in meta-analysis and is consistent with recommendations by the APA's Task Force on Statistical Inference.

Use of effect sizes promotes scientific inquiry because when a particular experimental study has been replicated, the different effect size estimates from those studies can be easily combined to produce an overall best estimate of the size of the intervention effect. The process of synthesizing results of experimental studies into an overall effect size is called meta-analysis. Meta-analysis assigns greater weight to larger sample studies (which are more precise) and lesser weight to smaller sample studies (which are less precise) to produce a weighted average effect size and its confidence interval. Meta-analysis also includes examining variability in effect sizes to determine whether the average effect size is representative of the effect sizes from which it was derived. When it is, clinicians should seriously consider the use of the intervention in their practice.

REFERENCES

Allison, P. D. (2001). *Missing data.* (Sage University Papers Series on Quantitative Applications in the Social Sciences, Series No. 07-136). Thousand Oaks, CA: Sage.

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

American Statistical Association. (1992). C*ombining information: Statistical issues and opportunities for research* (Contemporary Statistics, No. 1). Washington DC: National Academy.

Borenstein, M. (2005). Software for publication bias. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 193–220). West Sussex, England: Wiley.

Borenstein, M., & Rothstein, H. (1999). Comprehensive meta-analysis (Version 1.0) [Computer software]. Englewood, NJ: Biostat.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

Boudreau, L. A., & Jeffrey, C. J. (1973). Stuttering treated by desensitization. *Journal of Behavior Therapy and Experimental Psychiatry, 4,* 209–212.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Chalmers, I., & Altman, D. G. (1995). *Systematic reviews.* London: BMJ.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*(1), 12–37.

Clarke, M. (2002). The Cochrane Collaboration: Providing and obtaining the best evidence about the effects of health care. *Evaluation and the Health Professions, 25*(1), 8–11.

Coe, R. (2000). *What is an "effect size"? A brief introduction.* Unpublished manuscript, Durham University, UK.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cooper, H. (1998). *Synthesizing research.* Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1998). *Fundamentals of clinical trials.* New York: Springer.

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

Harris, V., Onslow, M., Packman, A., Harrison, E., & Menzies, R. (2002). An experimental investigation of the impact of the Lidcombe program on early stuttering. *Journal of Fluency Disorders, 27,* 203–214.

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

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis.* [ERIC Document Reproduction Service No. ED309952].

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327,* 557–560.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage.

Hunter, J. E., & Schmidt, F. L. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

James, J. E. (1976). The influence of duration on the effects of time-out from speaking. *Journal of Speech and Hearing Research, 19,* 206–215.

Jones, M., Onslow, M., Packman, A., Williams, S., Ormond, T., Schwarz, I., & Gebski, V. (2005). Randomised controlled trial of the Lidcombe programme of early stuttering intervention. *British Medical Journal, 331,* 659–661.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Last, J. M. (2001). *A dictionary of epidemiology*. Oxford, England: Oxford University Press

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Öst, L., Götestam, K. G., & Melin, L. (1976). A controlled study of two behavioral methods in the treatment of stuttering. *Behavior Therapy, 7,* 587–592.

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation & the Health Professions, 24*(3), 277–307.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis.* West Sussex, England: Wiley.

Ryan, B. P., & Van Kirk Ryan, B. (1995). Programmed stuttering treatment for children: Comparison of two establishment programs through transfer, maintenance, and follow-up. *Journal of Speech and Hearing Research, 38,* 61–75.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review, 11,* 157–169.

Waterloo, K. K., & Götestam, K. G. (1988). The regulated-breathing method for stuttering: An experimental evaluation. *Journal of Behavior Therapy and Experimental Psychiatry, 19,* 11–19.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Newbury Park, CA: Sage.

Contact author: Herbert M. Turner, III, PhD, 3700 Walnut Street, Philadelphia, PA 19104. E-mail: hmturner@gse.upenn.edu

APPENDIX A: WEB SITES FOR EFFECT SIZE CALCULATIONS AND META-ANALYSIS

The Cochrane Collaboration	www.cochrane.org
The Campbell Collaboration	www.campbellcollaboration.org
The What Works Clearinghouse	www.w-w-c.org
University of Murcia Meta-Analysis Unit	www.um.es/facpsi/metaanalysis
The Centre for Reviews and Dissemination (at York)	www.york.ac.uk/inst/crd
The EPPI Centre	http://eppi.ioe.ac.uk/EPPIWeb/home.aspx
Resources for Meta-Analysis	http://faculty.ucmerced.edu/wshadish/Meta-Analysis%20Links.htm

Additional Resources for Meta-Analysis at http://mason.gmu.edu/~dwilsonb/ma.html

APPENDIX B: SOFTWARE FOR EFFECT SIZE CALCULATIONS AND META-ANALYSIS

ES Effect Size Calculator	www.assess.com
Comprehensive Meta Analysis (CMA 2.0)	www.biostat.com
Stata (8.2)	www.stata.com
Review Manager (3.2)	www.cochranecollaboration.org
MetaWin (2.0)	www.metawinsoft.com