Interpreting and Maintaining the Evidence

Chad Nye
Jennine Harvey
University of Central Florida, Orlando

The goal is to transform data into information, and information into insight.

—Carly Fiorina
Former Chairman and CEO, Hewlett-Packard Co.

The work of systematic review, meta-analysis, and their application to evidence-based practice (EBP) is in fact the process of “transform[ing] data into information” and ultimately using that information to improve the quality of intervention for individuals with communication disorders. A systematic review and meta-analysis produces a summary of intervention outcomes described as measures of effect, efficacy, or effectiveness and defined as:

- **Effect**: the measurement of an intervention outcome(s) within a single study
- **Efficacy**: the measurement of common intervention outcomes across multiple studies, participant groups, methods, and contexts
- **Effectiveness**: the comparison of two or more known efficacious interventions

The preceding articles in this issue have provided a basic framework for understanding the process of conducting a systematic review and meta-analysis in such a way that the resulting data can be interpreted for practical application to best practice based on the available evidence. It is important to remember that this practical application will at least in part be determined by (a) how the data are interpreted, (b) who does the interpreting, and (c) the intended purposes of the interpretation. The data synthesis described in the previous article is the quantitative summary of the effects obtained for individual studies and the aggregation of those effects across multiple studies, comparison, or outcomes.

The purpose of this article is to describe how the effect size data that are produced in a systematic review can be interpreted to gain insight into the nature of effective intervention as well as some cautions in interpreting the effect size. In addition, we will suggest an ongoing procedure for maintaining systematic reviews for improved clinical service.

**Interpreting Effect Sizes**

**Magnitude of Effect**

At the most basic level of interpretation, the effect size of an intervention is a measure of the magnitude of impact...
resulting from an intervention. The larger the effect size, the greater the impact of the intervention. Because the effect size metric is a continuous metric that has both positive and negative directions, the magnitude is always set against the standard of 0.0, representing no effect. Thus, as the effect size moves to the positive side of the scale, the result is interpreted as being more beneficial to the intervention group than to the control (comparison) group. Similarly, as the effect size moves to the negative side of the scale, the result is interpreted as being more beneficial to the control group who did not receive the intervention; that is, the nontreated participants actually got better as a result of not participating in the intervention.

An important characteristic of the effect size calculation is that in situations in which the measured outcome for the intervention group is smaller than the measured outcome for the control group, the intervention result could still be a positive effect. For instance, when an intervention for persons who stutter reduces the number of stuttered syllables (i.e., the frequency of stuttering decreases), there is a positive effect and not a negative effect. The intervention appears to have a negative impact due to the direction sign of the data (e.g., −.33). However, the resulting effect size would be expressed as a positive impact on the treated group due to the conversion from negative to positive (i.e., reduced number of stuttered syllables), likewise if the conversion is reversed (i.e., increased number of stuttered syllables). In cases like this, the valence of the effect size would have to be reversed to make it interpretable in light of other effect sizes that have the opposite meaning.

So how do we know what the qualitative value is for increments of the effect size? How big is big? How small is small? The most frequently cited scale was put forth by Cohen (1988), in which he suggested that any effect size could be qualitatively categorized according to the following:

- Large ES = ≥ .80
- Moderate = .50
- Small = ≤ .20

With this scale, we can see that for effect sizes greater than .80 and under .20, it is a straightforward qualitative interpretation of the magnitude of effect—greater than .80 is large and the intervention was very effective, and less than .20 is small and the intervention made little difference in the postintervention measure. However, notice that that leaves a substantial range for which we only have the term “moderate” to capture the qualitative magnitude of effect. In general, we simply qualify the moderate with terminology such as “moderately high” or “low moderate.” The problem here is that the descriptive terms do not adequately reflect the precision we would expect given the calculation of a numeric effect size statistic, but it does give us a conceptual reference to describe the nature of the intervention effect.

There is another way to interpret the magnitude of outcome that expresses the intervention effect in terms of the percentage of treated participants improving as a result of the intervention. For example, if a comparison between the intervention and control groups resulted in an effect size of .85, we could consult a normal distribution table and locate the effect size value (treating the effect size as a z score) and the associated area covered under the normal distribution curve. Figure 1 illustrates a situation in which an effect size of .85 indicates that 80% of the intervention participants could, on average, be expected to improve by .85 SD as a result of the treatment. It is assumed that at the pretest level of random assignment, the mean of both groups was the same. Therefore, the change from the presumed no-effect (ES = 0.0) (50% statistical chance level) to the 80% level demonstrates a 30% improvement beyond chance for the intervention group. In other words, the effect size expresses an improvement that was 30% beyond what the expected progress would be if left to chance or with no intervention. Similarly, the remaining 20% of the treated participants would not have performed any better than the nontreated participants, as shown in Figure 1.

The problem with this interpretation is that we are still unable to make a definitive statement as to how much improvement occurred for the 80% of individuals whose postintervention performance improved by .85 SD. If this were the finding from a stuttering intervention, how many more nonstuttered words or how many fewer stuttered words would an effect size of .85 represent? Without a context of standardization by which to compare the intervention effect, we cannot convert the effect size into a more practice-based metric.

However, in some cases, it may be possible to convert the effect size into units of the outcome measurement that produced the effect size value (Lipsey & Wilson, 2001). With this conversion, the clinician could use the effect size to determine how much the intervention would change the score of the original outcome measure taken at pretreatment. For instance, if outcomes were measured using a standardized instrument in which the mean was 100 and the standard deviation was 10, then an effect size of .85 could be multiplied by the standard deviation to give the unit equivalent of 9, which when added to the mean of 100, would tell us that the intervention could, on average, result in an improvement of the measured outcome of 9 points to an average of 109. However, it is rare that the data for a systematic review and meta-analysis provide a consistent outcome measure across all studies. Even so, this procedure can be used with a subset of studies for which benchmark means and standard deviations are available.

---

**Figure 1.** Normal distribution curve with ES = .85.
In general, the basic interpretation of the effect size provides a relative standard reflecting the magnitude of the intervention effect. That is, the effect size tells us how much behavior change occurred when the intervention group was compared to the control group. However, it often does not tell us how much the behavior improved relative to a known standard of performance.

**Confidence Interval**

In spite of the difficulties of making practical sense of the effect size value as a single metric of effect, it does at least give us a starting point. But, the starting point is just that—the beginning. We also need to remember that the effect size is an *estimate* of effect and that that estimate is not absolute. That is, the effect size is a “best guess” given the available data. We know that in the behavioral sciences, our measurement is not as precise as we might like it to be, and the notion of *error* has to be considered in interpreting the observed effect size. To do this, we calculate a *confidence interval* (CI) around the effect size that helps us understand the precision of our estimate and, in fact, the potential generalization of the impact of the intervention. Without the CI, we only know that a given effect size is a measure of performance at one point in time, but we are also interested in knowing what we might expect if we use the intervention with other individuals with similar communication disorders.

In meta-analysis, we typically reference the effect size (sometimes referred to as the *point estimate*) with a 95% CI. The 95% CI is a statement of the potential range of effects if the intervention were given to other individuals in the population. The 95% CI implies that although a postintervention comparison provides a snapshot view of the average performance of the participants as a group, the possibility exists that there were some confounding features of the intervention that might not be present if we were to conduct the study again. So, the 95% CI represents a range of possible scores that we can accept with a fairly high level of confidence as being achievable on replication of the study. Sometimes, the effect size is referred to as the *observed effect*, and the 95% CI indicates the range of scores that contain the *true effect* at the 95% level of confidence—meaning that 95 times out of 100 we are confident that the interval contains the true effect in the population.

To illustrate the use and importance of the CI, Figure 2 presents a summary of a meta-analysis using what is called a forest plot (previously described in Turner and Bernard, 2006).

The values in the “Hedges’s g” column are the calculated effect sizes for each study that are weighted for sample size (ES†). The “Lower Limit” and “Upper Limit” columns represent the lower and upper limits of the 95% CI and provide us with a picture of the potential variability of the observed ES† of .854. Notice that the lower limit of the effect size is .507 and the upper limit is 1.202. With these pieces of data, we can say that we are 95% confident that the true effect of intervention for persons who stutter is no smaller than ES† = .507 and probably not any larger than ES† = 1.202. This gives us at least two very important pieces of information. First, because the lower limit effect size does not include 0.0, we can say that the effects are statistically significant; that is, the treated participants really improved because of the intervention. Note that the *p* value in the right column indicates the same result but presented as a *p* value. A “cause and effect” relationship can be accepted because (a) participants were randomly assigned to groups, and (b) the observed effect size was statistically significant. Second, we can say that the average intervention effect is at worst moderate and at best highly effective. That is, on a global level, intervention for

**Figure 2.** Overall effect for behavioral stuttering intervention: Treatment vs. control.
persons who stutter results in a positive improvement of speech behaviors. What we cannot say is that the same level of effect occurred for all studies. It is important to remember that the overall effect size does not represent the effect of any one single study.

In Figure 2, notice that of the six studies, at least five different types of intervention were identified. Clearly, the study by James (1976) produced virtually no effect on the participants’ stuttering, whereas Harris (2002) and Jones (2005) used the same intervention program and reported statistically significant and moderate to large intervention effects. So, does this mean that we can now eliminate any study that does not demonstrate statistically significant and large intervention effects? No, we cannot eliminate a single study because of a “poor showing,” just as we would not want to embrace a single study because of a “strong showing.” A single study with weak or nonsignificant results may have components that with minor modifications would result in an effective outcome. Similarly, a study with significant results may be the basis of further investigation into what components of the intervention may be driving the positive results. Remember, we do not make a diagnosis based on a single test; neither should we accept without question the efficacy of an intervention approach on the basis of a single study. Let us think through the rationale for this caution in rejecting or accepting the results of a single study in terms of the methodological quality of the individual study.

Methodological Quality

The methodological quality of the research is also a critical element of the evidence-based practice (EBP) decision (Moher, 1998). How are we to use individual studies, studies of weak scientific merit, or nonexperimental research in the decision-making process? Methodological quality refers to the degree to which a study has adhered to basic scientific principles in order to measure the outcomes of interest. Consider what has gone into the participant selection, intervention structure, and design procedures to produce the results of a study using the studies in Figure 2. For example, the James (1976) study produced a nonsignificant and small intervention effect but the intervention was provided to adults on a single-session basis. On the other hand, the Harris, Onslow, Packman, Harrison, and Menzies (2002) and Jones, Onslow, Packman, Williams, Ormond, Schwarz, and Gebski (2005) studies reported moderate to large intervention effects. In both of these studies, the intervention was conducted over a 12-week period and was directed at parents of children who stuttered. Was the difference due to one of the participant or intervention dimensions? Could it be that the small effect size in the James study was due to the short duration of the intervention? Perhaps the Harris and Jones studies produced such large effects because the intervention period was longer and the parents were trained to serve as a more powerful discriminative stimulus for fluent speech for a longer period of the child’s day?

Questions regarding participant or intervention characteristics are the kinds of issues that the researcher would want to consider if a research agenda was developed to study either the time-out (TO) intervention used by James or the Lidcombe parent training program presented by Harris and Jones. From an EBP point of view, the clinician might be more inclined to implement the Lidcombe program because there is more available evidence from studies using strong scientific standards of investigation. At the very least, further evidence is needed to draw defensible conclusions for intervention guidelines if TO is to be the intervention of choice.

The effect size, whether weighted (ES+) or unweighted (ES), and associated 95% CI provides the clinician with a quantitative standard by which decisions for clinical practice can be adopted. The way the study is structured can have an impact on the magnitude to the intervention effect size. There are numerous variables in each study that can confound the interpretation of the observed intervention. A portfolio of evidence consisting of multiple studies is necessary to support or reject the effectiveness of an intervention that is to be used to guide clinical decisions regarding the choice of intervention.

Maintaining Evidence

The decision to reject or adopt a specific intervention is not the sole purpose of a systematic review, meta-analysis, or the effect size data. The collection, summary, analysis, and interpretation of intervention effects are all important and necessary functions of developing an evidence-based portfolio of research that can guide the practice of speech-language pathologists. It would, however, be a mistake to think of this as a one-shot solution to meeting the need for EBP. Current research and practice should be a cumulative process of building a knowledge base from past research. In order to maintain a viable and useful EBP, a commitment also has to be made to continually update the portfolio. The need for a review of systematic reviews and meta-analyses on a regular basis will be critical to improving the delivery of clinical services as well as the communication skills of individuals being treated for communication disorders. As new research is made available, those data need to be included in existing reviews and analyses and the results reinterpreted if necessary. It is certainly possible that with more evidence in a given area, the strength and direction of intervention could be altered.

The process of transforming data to information and information to insight means that intervention data have to be understood in the context of data, clinical expertise, and client satisfaction. A systematic review can open windows of insight as to what works and what does not even if the evidence is thin or weak. Nobel Laureate, Albert Szent-Gyorgyi (n.d.), said it well, “Discovery consists of seeing what everybody has seen and thinking what nobody has thought.”
REFERENCES


Contact author: Chad Nye, PhD, Executive Director, University of Central Florida, Center for Autism & Related Disabilities, 12001 Science Dr., Suite 145, Orlando, FL 32826. E-mail: cnye@mail.ucf.edu