

Development, Procedures, and Application of the Evaluative Method for Determining Evidence-Based Practices in Autism

Brian Reichow

ABBREVIATIONS

ASDs	Autism spectrum disorders
EBP	Evidence-based practice
SSED	Single subject experimental design

BACKGROUND

Recently, we sought to review the empirical evidence on interventions for children with autism spectrum disorders (ASDs) in search of interventions meeting the criteria of evidence-based practice (EBP). As outlined in Chap. 1, EBP is defined differently by different disciplines. Although many of these definitions were quite good, it quickly became apparent that locating and defining EBP for children with ASDs using the available definitions and procedures would be difficult. Therefore, we decided to create a new

method for evaluating empirical evidence to determine if a practice could be considered an EBP. This decision was made only after determining that existing methods were not well suited for our specific needs; when possible, elements and standards from existing methods were adopted into the evaluative method presented in this chapter.

The evaluative method (Reichow et al. 2008) was created to assist with the identification of practices that could be considered EBPs for children with ASDs. It provides a method of evaluating intervention research and includes three instruments: rubrics for the evaluation of research report rigor; guidelines for the evaluation of research report strength; and criteria for determining if an intervention has the evidence needed to be considered an EBP. Initial assessments suggest that the evaluative method is a tool that can be used reliably to review intervention research to produce valid assessments of the empirical evidence on practices for children with ASDs (see Chap. 3).

This chapter expands upon previous presentations of the evaluative method in two ways. First, it provides operationalized definitions and rating criteria for the primary and secondary quality indicators for the rubrics. Secondly, it provides a formula (i.e., an algorithm) that can be used to apply the EBP criteria to a set of studies to determine the EBP status of an intervention.

RESEARCH REPORT RIGOR

To evaluate the rigor of research reports, two rubrics were developed, one for research conducted using group research methods and one for research conducted using single subject experimental designs (SSED). These rubrics provide a grading scheme that evaluates the quality (i.e., the rigor) of methodological elements of individual research reports. Two levels of methodological element are included in the rubrics (Table 2.1): primary quality indicators and secondary quality indicators. Primary quality indicators are elements of the research methodology deemed critical for demonstrating the validity of a study. They are operationally defined and graded on a trichotomous ordinal scale (high quality, acceptable quality, and unacceptable quality). The secondary quality indicators are elements of research design that, although important, are not deemed necessary for the establishment of the validity of a study. They are operationally defined on a dichotomous scale (the report either contains or does not contain evidence of each indicator).

Because high-integrity experiments of group research designs and SSED share many characteristics, attempts were undertaken to retain similar definitions across rubrics. However, indicators specific to one type of research method are needed due to the differences in research methodologies (see Table 2.1).

When using the rubrics to evaluate research reports, we have found it helpful to create separate scoring sheets for each type of research methodology. Examples

TABLE 2.1 Primary and secondary quality indicators by type of experimental design

Group research designs	Single subject experimental design
Primary quality indicators	
• Participant characteristics	• Participant characteristics
• Independent variable	• Independent variable
• Comparison condition	• Dependent variable
• Dependent variable	• Baseline condition
• Link between research question and data analysis	• Visual analysis
• Statistical analysis	• Experimental control
Secondary quality indicators	
• Random assignment	• Interobserver agreement
• Interobserver agreement	• Kappa
• Blind raters	• Blind raters
• Fidelity	• Fidelity
• Attrition	• Generalization or maintenance
• Generalization or maintenance	• Social validity
• Effect size	
• Social validity	

of these scoring sheets are shown in [Appendices 1](#) and [2](#) for group research designs and SSED, respectively. The following sections discuss the criteria for allocating a rating under each indicator.

Primary Quality Indicators for Group Research Design

Participant characteristics (PART) A high (H) quality rating is awarded to a study that meets the following criteria:

- 1. Age and gender are provided for all participants (mean age is acceptable).

2. All participants' diagnoses are operationalized by including the specific diagnosis and diagnostic instrument (acceptable instruments include ADOS, ADI-R, CARS, DSM-IV, and ICD-10) used to make the diagnosis or an operational definition of behaviors and symptoms of the participants.
3. Information on the characteristics of the interventionist are provided (the ability to determine who did the intervention is minimal a criterion) and information on any secondary participants (e.g., peers) is provided.
4. If a study provides standardized test scores, the measures used to obtain those scores are indicated.

An acceptable (A) quality rating is awarded to a study that meets criteria 1, 3 and 4. A study that does not meet all of criteria 1, 3, and 4 is of unacceptable quality and is awarded a U rating.

Independent variable (IV) (e.g., *intervention*) An H rating is awarded to a study that defines independent variables with replicable precision (i.e., one could reproduce the intervention given the description provided). If a manual is used, the study passes this criterion. An A rating is awarded to a study that defines many elements of the independent variable but omits specific details. A U rating is awarded to a study that does not sufficiently define the independent variables.

Comparison condition (CC) An H rating is awarded to a study that defines the conditions for the comparison group with replicable precision, including a description of any other interventions participants receive. An A rating is awarded to a study that vaguely describes the conditions for the comparison group; information on other interventions may not be reported. A U rating is awarded to a study that does not report the conditions for the comparison group or has no control or comparison group.

Dependent variable (DV) or outcome measure An H rating is awarded to a study that meets the following criteria:

- The variables are defined with operational precision.
- The details necessary to replicate the measures are provided.
- The measures are linked to the dependent variables.
- The measurement data is collected at appropriate times during the study for the analysis being conducted.

An A rating is awarded to a study that meets three of the four criteria. A U rating is awarded to a study that meets fewer criteria.

Link between research question and data analysis (LRQ) An H rating is awarded to a study in which data analysis is strongly linked to the research questions and uses correct units of measure (i.e., child level, teacher level, etc.) on all variables. An A rating is awarded to a study in which data analysis is poorly linked to the research questions but uses correct units for a majority of the outcome measures. A U rating is awarded to a study in which data analysis is linked weakly or not at all to the research questions and uses the correct unit for only a minority of the outcome measures.

Statistical analysis (STAT) An H rating is awarded to a study in which proper statistical analyses were conducted with an adequate power and sample size ($n > 10$) for each statistical measure. An A rating is awarded to a study in which proper statistical analyses were conducted for at least 75% of the outcome measures or in which proper statistical analyses were conducted on 100% of outcome measures but with inadequate power or a small sample size. A U rating is awarded to a study in which statistical analysis was not done correctly, the sample size was too small or the power was inadequate.

Secondary Quality Indicators for Group Research Design

These indicators are rated on a dichotomous scale (there either is, or is not, evidence of the indicator).

Random Assignment (RA) This indicator is positive if participants are assigned to groups using a random assignment procedure.

Interobserver Agreement (IOA) This indicator is positive if IOA is collected across all conditions, raters, and participants with reliability $>.80$ (Kappa $>.60$) or psychometric properties of standardized tests are reported and are $>.70$ agreement with a Kappa $>.40$.

Blind Raters (BR) This indicator is positive if raters are blind to the treatment condition of the participants.

Fidelity (FID) This indicator is positive if treatment or procedural fidelity is continuously assessed across participants, conditions, and implementers, and if applicable, has measurement statistics $>.80$.

Attrition (ATR) This indicator is positive if articulation is comparable (does not differ between groups by more than 25%) across conditions and less than 30% at the final outcome measure.

Generalization or Maintenance (G/M) This indicator is positive if outcome measures are collected after the final data collection to assess generalization or maintenance.

Effect Size (ES) This indicator is positive if effect sizes are reported for at least 75% of the outcome measures and are $>.40$.

Social Validity (SV) This indicator is positive if the study contains at least four of the following features:

- Socially important DVs (i.e., society would value the changes in outcome of the study)
- Time- and cost-effective intervention (i.e., the ends justify the means)
- Comparisons between individuals with and without disabilities
- A behavioral change that is large enough for practical value (i.e., it is clinically significant)

- Consumers who are satisfied with the results
- IV manipulation by people who typically come into contact with the participant
- A natural context

Primary Quality Indicators for SSEs

Participant Characteristics (PART) A high (H) quality rating is awarded to a study that meets the following criteria:

1. Age and gender are provided for all participants.
2. All participants' diagnoses are operationalized by including the specific diagnosis and diagnostic instrument (acceptable instruments include ADOS, ADI-R, CARS, DSM-IV, and ICD-10) used to make the diagnosis or an operational definition of behaviors and symptoms of the participants.
3. Information on the characteristics of the interventionist are provided (the ability to determine who did the intervention is a minimal criterion) and information on any secondary participants (e.g., peers) is provided.
4. If a study provides standardized test scores, the measures used to obtain those scores are indicated.

An acceptable (A) quality rating is awarded to a study that meets criteria 1, 3, and 4. A study that does not meet all of criteria 1, 3, and 4 is of unacceptable quality and is awarded a U rating.

Independent Variable (IV) (e.g., *intervention*) An H rating is awarded to a study that defines independent variables with replicable precision (i.e., you could reproduce the intervention given the description provided). If a manual is used, the study passes this criterion. An A rating is awarded to a study that defines many elements of the independent variable but omits specific details. A U rating is awarded to a study

that does not sufficiently define the independent variables.

Baseline Condition (BSLN) An H rating is awarded to a study in which 100% of baselines:

- Encompass at least three measurement points
- Appear through visual analysis to be stable
- Have no trend or a counter-therapeutic trend
- Have conditions that are operationally defined with replicable precision

An A rating is awarded to a study in which at least one of the above criteria was not met in at least one, but not more than 50%, of the baselines. A U rating is awarded to a study in which two or more of the above criteria were not met in at least one baseline or more than 50% of the baselines do not meet three of the criteria.

Dependent variable (DV) or outcome measure An H rating is awarded to a study that meets the following criteria:

- The variables are defined with operational precision.
- The details necessary to replicate the measures are provided.
- The measures are linked to the dependent variables.
- The measurement data is collected at appropriate times during the study for the analysis being conducted.

An A rating is awarded to a study that meets three of the four criteria. A U rating is awarded to a study that meets fewer criteria.

Visual Analysis (VIS ANAL) An H rating is awarded to a study in which 100% of graphs (i.e., tiers within a figure):

- Have data that are stable (level or trend)
- Contain less than 25% overlap of data points between adjacent conditions, unless behavior is at ceiling or floor levels in the previous condition

- Show a large shift in level or trend between adjacent conditions that coincide with the implementation or removal of the IV. If there was a delay in change at the manipulation of the IV, the study is accepted as high quality if the delay was similar across different conditions or participants (+/-50% of delay)

An A rating is awarded to a study in which two of the criteria were met on at least 66% of the graphs. A U rating is awarded to a study in which two or fewer criteria were met on less than 66% of the graphs.

Experimental Control (EXP CON) An H rating is awarded to a study that contains at least three demonstrations of the experimental effect, occurring at three different points in time and changes in the DVs vary with the manipulation of the IV in all instances of replication. If there was a delay in change at the manipulation of the IV, the study is accepted as high quality if the delay was similar across different conditions or participants (+/-50% of delay). An A rating is awarded to a study in which at least 50% of the demonstrations of the experimental effect meet the above criteria, there are two demonstrations of the experimental effect at two different points in time and changes in the DVs vary with the manipulation of the IV. A U rating is awarded to a study in which less than 50% of the demonstrations of the experimental effect meet the above criteria, there are fewer than two demonstrations of the experimental effect occurring at two different points in which changes in the DVs vary with the manipulation of the IV.

Secondary Quality Indicators for SSEDs

These indicators are rated on a dichotomous scale (there either is, or is not, evidence of the indicator).

Interobserver Agreement (IOA) This indicator is positive if IOA is collected across all conditions, raters, and participants with reliability >.80.

Kappa (KAP) This indicator is positive if Kappa is calculated on at least 20% of sessions across all conditions, raters, and participants with a score $>.60$.

Blind Raters (BR) This indicator is positive if raters are blind to the treatment condition of the participants.

Fidelity (FID) This indicator is positive if treatment or procedural fidelity is continuously assessed across participants, conditions, and implementers, and if applicable, has measurement statistics $>.80$.

Generalization or Maintenance (G/M) This indicator is positive if outcome measures are collected after the final data collection to assess generalization or maintenance.

Social Validity (SV) This indicator is positive if the study contains at least four of the following features:

- Socially important DVs (i.e., society would value the changes in outcome of the study)
- Time- and cost-effective intervention (i.e., the ends justify the means)
- Comparisons between individuals with and without disabilities

- A behavioral change that is large enough for practical value (i.e., it is clinically significant)
- Consumers who are satisfied with the results
- IV manipulation by people who typically come into contact with the participant
- A natural context

RESEARCH REPORT STRENGTH RATINGS

The second instrument of the evaluative method provides scoring criteria to synthesize the ratings from the rubrics into a rating of the strength of the research report. There are three levels of research report strength: strong, adequate, and weak. The requirements for each strength rating are shown in Table 2.2. Research reports with a strong rating demonstrate concrete evidence of high quality. These reports received high quality grades on all primary indicators and contained evidence of multiple secondary quality indicators.

TABLE 2.2 Guidelines for the determination of research report strength ratings (Adapted from Reichow et al. 2008. With permission)

Strength rating	Group research	Single subject research
Strong	Received high quality grades on all primary quality indicators and showed evidence of four or more secondary quality indicators	Received high quality grades on all primary quality indicators and showed evidence of three or more secondary quality indicators
Adequate	Received high quality grades on four or more primary quality indicators with no unacceptable quality grades on any primary quality indicators, and showed evidence of at least two secondary quality indicators	Received high quality grades on four or more primary quality indicators with no unacceptable quality grades on any primary quality indicators, and showed evidence of at least two secondary quality indicators
Weak	Received fewer than four high quality grades on primary quality indicators or showed evidence of less than two secondary quality indicators	Received fewer than four high quality grades on primary quality indicators or showed evidence of less than two secondary quality indicators

An adequate rating designates research showing strong evidence in most, but not all areas. Elements of reports achieving an adequate rating might have received acceptable grades on up to two primary quality indicators and must have shown evidence of at least two secondary quality indicators. A study receiving an adequate strength rating cannot receive an unacceptable grade on any primary quality indicator. A weak rating indicates that the research report has many missing elements or fatal flaws. Reports receiving one or more unacceptable grades on primary quality indicators or evidence of one or fewer secondary quality indicators receive a weak rating. Because conclusions about the results of a

study receiving a weak rating are tentative at best, studies receiving this rating are not used when determining the EBP status of an intervention.

CRITERIA FOR LEVELS OF EBP

The final instrument provides the criteria for the aggregation of research reports with respect to their strength rating across studies to determine whether a practice has amassed enough empirical support to be classified as an EBP. The criteria for two levels of EBP, established and promising (see Table 2.3), have been guided by

TABLE 2.3 Criteria for treatments to be considered EBP (Adapted from Reichow et al. 2008. With permission)

Level of EBP	Example criteria
Established (≥ 60 points from the EBP status formula)	<ul style="list-style-type: none"> • Five SSED studies of strong research report strength with a total sample size of at least 15 participants across studies conducted by at least three research teams in three different geographic locations • Ten SSED studies of adequate research report strength with a total sample size of at least 30 different participants across studies conducted by at least three research teams in three different geographic locations • Two group design studies of strong research report strength conducted by in different geographic locations • Four group design studies of at least adequate research report strength conducted in at least two different research teams • One group design study of strong research report strength and three SSED studies of strong research report strength with at least 8 different participants • Two group design studies of at least adequate research report strength and six SSED studies of at least adequate research report strength with at least 16 different participants
Promising (> 30 points from the EBP status formula)	<ul style="list-style-type: none"> • Five SSED studies of at least adequate research report strength with a total sample size of at least 16 different participants across studies conducted by at least two research teams in two different geographic locations • Two group design studies of at least adequate research report strength • One group research report of at least adequate research report strength rating and at least three SSED studies of at least adequate strength rating with at least 8 participants

previous EBP criteria (Filipek et al. 2000; Gersten et al. 2005; Horner et al. 2005; Kratochwill and Stoiber 2002; Lonigan et al. 1998) and contain an operationalization for documenting the evidence needed to meet the criteria of the two levels. A treatment must meet at least one example criterion; it can meet multiple criteria.

An established EBP is a treatment shown to be effective across multiple methodologically sound studies conducted by at least two independent research groups in separate geographical locations. Practices meeting this requirement have demonstrated enough evidence for confidence in the treatment's efficacy. A promising EBP is also a treatment shown to be effective across multiple studies but for which the evidence is limited by weaker methodological rigor, fewer replications, and/or an inadequate number of independent researchers demonstrating the effects.

The two levels of EBP status can be obtained only using studies conducted using group research designs, only using studies conducted using SSED, or by using a combination of studies conducted using group research designs and SSED. When determining the EBP status of an intervention in which the evidence was obtained using group research designs, it is necessary to look at how many studies have been conducted. When determining the EBP status of an intervention in which the evidence was obtained using SSED, it is necessary to examine both the number of studies that were conducted and the number of participants with whom the procedures have been replicated. When examining the number of participants, an adaptation of the success estimate created by Reichow and Volkmar (*in press*) is recommended. The adapted success estimate should be estimated using visual analysis (Gast and Spriggs 2010); it provides an estimate of the number of participants for whom the intervention was successful within each study. Using the adapted success estimate

should provide a more accurate appraisal of the number of participants for whom the intervention has worked than using the total number of participants from a study, as has been suggested in previous definitions of EBP (Horner et al. 2005; Lonigan et al. 1998).

The EBP status formula (2.1) provides a tool that can be used to assess all possible combinations of evidence that can be pooled to demonstrate the efficacy of a practice with respect to its status as an EBP.

$$\begin{aligned} &(\text{Group}_s * 30) + (\text{Group}_A * 15) + \\ &(\text{SSED}_s * 4) + (\text{SSED}_A * 2) = Z \quad (2.1) \end{aligned}$$

Group_s is the number of studies conducted using group research designs earning a strong rating, Group_A is the number of studies conducted using group research designs earning an adequate rating, SSED_s is the number of participants for whom the intervention was successful from SSED studies earning a strong rating, SSED_A is the number of participants for whom the intervention was successful from SSED studies earning an adequate rating, and Z is the total number of points for an intervention. It was determined that eight participants from strong SSED studies were equivalent to one strong group study by averaging two previous definitions of EBP providing a quantification of the number of SSED participants needed to achieve the highest level of evidence (Horner et al. 2005; Lonigan et al. 1998). When using the formula, 31 points are required for an intervention to meet the criteria of a promising EBP and 60 points are required for an intervention to meet the criteria of an established EBP. The criterion point levels are set such that there must be at least two studies for a practice to meet either EBP criterion and the formula is weighted such that studies with strong ratings contribute twice as much as studies receiving adequate rigor ratings. A reproducible worksheet containing blanks in

which to place the necessary information to calculate the EBP status formula is provided in [Appendix 3](#). Table 2.3 presents a sample, but not an exhaustive list, of ways in which the criteria for EBP can be met with studies conducted using only group research designs, using only SSED, and through combinations of the two research methodologies. Because this is an early attempt at synthesizing studies conducted using group research designs and SSED, empirical validation of the criteria is needed.

APPLICATION OF THE EBP STATUS FORMULA AND EBP CRITERIA

This section provides three examples of synthesizing multiple research report strength ratings for an intervention to determine whether the intervention has demonstrated the quantity and quality of

evidence to be considered an EBP. The first two examples are taken from a recent review of interventions for increasing prosocial behavior by Reichow and Volkmar ([in press](#)). Based on the results of their review, two interventions for school-aged children met the criteria of EBP.

Synthesizing Group Results

Table 2.4 provides an example of using the EBP status formula and applying the criteria of EBP using the results for social skills group interventions. As shown in the table, two studies using group research methodology received strong rigor ratings, thus $Group_s=2$. Using Formula (2.1), social skills groups amassed 60 points, which meets the level for an established EBP and so social skills groups for school-aged children can be classified as an established EBP. Note, because only group studies were used, the number of participants in each study was not applicable and did not affect the calculation.

TABLE 2.4 EBP status of social skills groups for school-aged children (As reviewed by Reichow and Volkmar [in press](#))

Study	Research method	Rigor rating	Successful <i>N</i>						
Lopata et al. (2008)	Group	Strong	N/A						
Owens et al. (2008)	Group	Strong	N/A						
Number of group <i>studies</i> with strong rigor ratings		2	= Group _s						
Number of group <i>studies</i> with adequate rigor ratings		0	= Group _A						
Number of <i>participants</i> from SSED studies with strong rigor ratings		0	= SSED _s						
Number of <i>participants</i> from SSED studies with adequate rigor ratings		0	= SSED _A						
Formula for determining EBP status									
$(Group_s * 30) + (Group_A * 15) + (SSED_s * 4) + (SSED_A * 2) = Z$									
$(2 * 30) + (0 * 15) + (0 * 4) + (0 * 2) = Z$									
$60 = Z$									
Points (<i>Z</i>)	0	10	20	30	31	40	50	59	60+
EBP status	Not an EBP				Probable EBP				Established EBP

Synthesizing SSED Results

The application of the results of the Reichow and Volkmar (in press) synthesis is presented for video modeling for school-aged children in Table 2.5. As shown, five studies using SSED methods received adequate rigor ratings. Because there were a total of 16 participants across the five studies, $SSED_A = 16$. Using Formula (2.1), video modeling amassed 32 points, which is below the criteria for either level of EBP. Thus, video modeling for school-aged children can be classified as a promising EBP. Note, in contrast to the first example, which used only group design studies, the second example used only SSED studies; thus, all calculations involved the number participants for whom the intervention was successful from each of the studies.

Synthesizing Group and SSED Results

An example of synthesizing results across studies conducted using both group research designs and SSED is illustrated from the analysis of behavioral interventions to improve joint attention behaviors by Ferraioli and Harris (see Chap. 6). As shown in Table 2.6, one group research study received a strong rigor rating ($Group_s = 1$). The remaining studies were conducted using SSED; four studies with 15 participants for whom the interventions were successful received strong rigor ratings ($SSED_s = 15$) and one study with two participants for whom the intervention was successful received an adequate rigor rating ($SSED_A = 2$). In summing the values, behavioral interventions for increasing joint attention behav-

TABLE 2.5 EBP status of video modeling for school-aged children (As reviewed by Reichow and Volkmar in press)

Study	Research method	Rigor rating	Successful N
Buggey (2005)	SSED	Adequate	2
Charlop-Christy and Daneshvar (2003)	SSED	Adequate	3
Nikopoulous and Keenan (2004)	SSED	Adequate	3
Nikopoulous and Keenan (2007)	SSED	Adequate	3
Sherer et al. (2001)	SSED	Adequate	5
Number of group <i>studies</i> with strong rigor ratings		0	= $Group_s$
Number of group <i>studies</i> with adequate rigor ratings		0	= $Group_A$
Number of <i>participants</i> from SSED studies with strong rigor ratings		0	= $SSED_s$
Number of <i>participants</i> from SSED studies with adequate rigor ratings		16	= $SSED_A$

Formula for determining EBP status

$$\begin{aligned}
 & (Group_s * 30) + (Group_A * 15) + (SSED_s * 4) + (SSED_A * 2) = Z \\
 & (0 * 30) + (0 * 15) + (0 * 4) + (16 * 2) = Z \\
 & 32 = Z
 \end{aligned}$$

Points (Z)	0	10	20	30	31	40	50	59	60+
EBP status	Not an EBP				Probable EBP				Established EBP

TABLE 2.6 EBP status of behavioral interventions to increase joint attention behaviors (As reviewed by Ferraioli and Harris in Chap. 6)

Study	Research method	Rigor rating	Successful <i>N</i>
Kasari et al. (2006)	Group	Strong	N/A
Martins and Harris (2006)	SSED	Strong	3
Rocha et al. (2007)	SSED	Strong	3
Whalen and Schreibman (2003)	SSED	Strong	5
Whalen et al. (2006)	SSED	Strong	4
Zercher et al. (2001)	SSED	Adequate	2
Number of group <i>studies</i> with strong rigor ratings		1	= Group _S
Number of group <i>studies</i> with adequate rigor ratings		0	= Group _A
Number of <i>participants</i> from SSED studies with strong rigor ratings		15	= SSED _S
Number of <i>participants</i> from SSED studies with adequate rigor ratings		2	= SSED _A

Formula for Determining EBP Status

$$\begin{aligned}
 &(\text{Group}_S * 30) + (\text{Group}_A * 15) + (\text{SSED}_S * 4) + (\text{SSED}_A * 2) = Z \\
 &(1 * 30) + (0 * 15) + (15 * 4) + (2 * 2) = Z \\
 &94 = Z
 \end{aligned}$$

Points (<i>Z</i>)	0	10	20	30	31	40	50	59	60+
EBP Status	Not an EBP				Probable EBP				Established EBP

iors amassed 94 points, which exceeds the criterion for an established EBP. Because the number of participants for whom the interventions were successful in each study was a factor in the calculations for the SSED variables, the number of participants did factor into the determination of EBP status.

DISCUSSION

In principle, the arrangement of identifying educational practices based on scientific evidence is admirable; using scientific evidence to inform practice should increase the likelihood of providing effective treatments. However, researchers have established few EBP for young children with ASDs through the application of established operationalized criteria (Reichow and Volkmar in Chap. 1;

Rogers and Vismara 2008). The evaluative method outlined in this chapter was created to address the need to identify EBP for young children with ASDs. The method has been used to examine the state of the science in research involving young children with ASDs (Reichow et al. 2007); to evaluate the empirical evidence on the Picture Exchange Communication System (Doehring et al. 2007); to determine the methodological rigor of studies included in research syntheses (Reichow and Volkmar in press; Reichow and Wolery 2009); and to determine the EBP status of interventions to increase the prosocial behavior of individuals with ASDs (Reichow and Volkmar in press). Collectively, the applications of the evaluative method have led to a number of practices being identified as EBP.

One particularly noteworthy characteristic of the criteria of EBP is the combination of multiple research methodologies.

The evaluative method of Reichow et al. (2008) was one of the first conceptualizations of EBP to provide an operationalized method for combining multiple research methods to investigate a single practice. The National Standards Project (National Autism Center 2009) described in Chap. 1 is a second example of synthesizing studies across research designs. Although the innovation of operationalizing a method of combining group research designs and SSED is noteworthy, the inclusion of only these research methodologies is limiting. However, our experience of reviewing the intervention literature for individuals with ASDs leads us to believe that the overwhelming majority of studies have been and continue to be conducted using the two methodologies for which the evaluative method was designed (i.e., group research designs and SSED). Thus, apart from a philosophical objection, it is unclear how much this limitation affects the results of a review conducted using the evaluative method.

The evaluative method has boundaries to its use and is not without limitations. With respect to the boundaries, the method was designed to evaluate research reports of specific interventions (e.g., focal interventions), not comprehensive programs. Second, the method was designed to evaluate individual experimental research reports; thus, the method is not appropriate for evaluating the methodological rigor of systematic reviews or meta-analyses. Since many hierarchies of evidence consider findings from multiple systematic reviews to be the highest level of evidence (Straus et al. 2005), this is a limitation that needs to be addressed in the future. Finally, some of the quality indicators on the rubrics are weighted toward studies demonstrating positive effects (e.g., effect size, visual analysis, and experimental control). Because the rubric for SSED contains these elements as primary quality indicators, a study

failing to demonstrate a positive effect would most likely receive an unacceptable grade on these elements and thus be rated as a weak study. Some definitions of EBP have created categories and grading schemes for treatments that do not work (Hawley and Wiesz 2002; Lilienfeld 2005) or for which there is conflicting evidence (De Los Reyes and Kazdin 2008). Making recommendations on ineffective practices was not a goal of the initial project and it is unlikely that the evaluative method in its current state is appropriate for evaluating research with null or negative results. Thus, the evaluative method is best suited for evaluating empirical research on interventions for individuals with ASDs conducted using group research designs or SSED in which the desired change in behavior was achieved.

Five additional limitations should be acknowledged. First, it is likely that there is a minimum set of competencies or knowledge base for using the evaluative method. Because the evaluative method is still relatively new and has not been widely tested, the extent of this limitation is not known. However, we can report that it has been successfully used and adopted by professionals not involved in this project (Children's Services Evidence-Based Practice Advisory Committee 2009; V. Smith, November 22, 2007 in a personal communication). Second, some elements of the evaluative method (e.g., the number of participants needed for a practice to be considered an EBP when using SSED) have not been empirically validated. Third, although the preliminary assessments of reliability and validity were positive (see Chap. 3), these assessments were conducted on a small sample by individuals closely connected with the development of the method. Validation of its application in real-world settings remains to be seen. Fourth, the evaluative method contains no method for assessing publication bias (Borenstein et al. 2009). Although the

inclusion of an instrument capable of synthesizing research across multiple research methodologies is likely to lessen the impact of publication bias when compared to other methods of determining EBP, the threat of publication bias remains and should be acknowledged. Finally, as with any review, application of the evaluative method provides a picture of the research that was reviewed in that instance. Different reviews occurring at different times with different inclusion criteria are likely to produce conflicting results.

REFERENCES

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: Wiley.
- Children's Services Evidence-Based Practice Advisory Committee. (2009). *Interventions for autism spectrum disorders: State of the evidence*. Portland.
- De Los Reyes, A., & Kazdin, A. E. (2008). When the evidence says, "yes, no, and maybe so": Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science*, 17, 47–51.
- Doehring, P., Reichow, B., & Volkmar, F. R. (2007). Is it evidenced-based? How to evaluate claims of effectiveness for autism. Paper presented at the International Association for Positive Behavior Support Conference, March, Boston, MA.
- Filipek, P. A., Accardo, P. J., Ashwal, S., Baranek, G. T., Cook, E. H., Jr., et al. (2000). Practice parameter: Screening and diagnosis of autism. *Neurology*, 55, 468–479.
- Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research in behavioral methodology* (pp. 199–233). New York: Routledge.
- Gersten, R., Fuchs, L. S., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, 71, 149–164.
- Hawley, K. M., & Weisz, J. R. (2002). Increasing the relevance of evidence-based treatments review to practitioners and consumers. *Clinical Psychology: Science and Practice*, 9, 225–230.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165–179.
- Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the procedural and coding manual of Division 16 and the Society for the Study of School Psychology task force. *School Psychology Quarterly*, 17, 341–389.
- Lilienfeld, S. O. (2005). Scientifically unsupported and supported interventions for childhood psychopathology: A summary. *Pediatrics*, 115, 761–764.
- Lonigan, C. J., Elbert, J. C., & Johnson, S. B. (1998). Empirically supported psychosocial interventions for children: An overview. *Journal of Clinical Child Psychology*, 27, 138–145.
- Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., et al. (2005). Challenges in evaluating psychosocial interventions for autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, 35, 695–711.
- National Autism Center. (2009). *National standards report*. Randolph: National Autism Center.
- Reichow, B., & Volkmar, F. R. (in press) Introduction to evidence-based practices in autism. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism*. New York, NY: Springer.
- Reichow, B., & Wolery, M. (2009). Comprehensive synthesis of early intensive behavioral interventions for young children with autism based on the UCLA Young Autism Project model. *Journal of Autism and Developmental Disorders*, 39, 23–41.
- Reichow, B., Barton, E. E., Volkmar, F. R., & Cicchetti, D. V. (2007). The status of research on interventions for young children with autism spectrum disorders. Poster presented at the International Meeting for Autism Research, May, Seattle, WA.
- Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (2008). Development of an evaluative method for determining the strength of research evidence in autism. *Journal of Autism and Developmental Disorders*, 38, 1311–1319.
- Rogers, S. J., & Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology*, 37, 8–38.
- Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). Edinburgh: Elsevier.

[illegible][illegible]

Evidence-Based Practices and Treatments for Children
with Autism

Reichow, B.; Doehring, P.; Cicchetti, D.V.; Volkmar, F.R.
(Eds.)

2011, XVIII, 408 p. 4 illus., Hardcover

ISBN: 978-1-4419-6973-6