The Quality of Single-Case Evaluation Studies of Curricular Programs for Students with Disabilities

Ronald C. Martella, Ph.D., BCBA-D
Purdue University
College of Education

J. Ron Nelson, Ph.D.
University of Nebraska

Nancy E. Marchand-Martella, Ph.D., BCBA-D
Purdue University

Tracy E. Sinclair, M.Ed.
University of Oklahoma

Abstract: *Educational researchers conduct studies to gather critical empirical evidence in the determination of what are "evidence-based" curricular programs, which in turn, directly inform adoption efforts. The predominate method of validating these programs is through the use of group experimental designs, although single-case designs have also been advocated. This article posits the current standards set forth when validating curricular programs using single-case designs are significantly lacking. We propose an expansion of the existing standards to ensure a rigorous, accurate examination of curricular programs when determining their use as an evidence-based practice. We apply these standards to a subset of studies reviewed by McKenna, Kim, Shin, and Pfannenstiel (2017) in a recent evaluation of single-case reading intervention investigations including students with and at risk for emotional and behavioral disorders. Using our expanded standards, we found none of the 14 studies met quality standards for curricular program validation. Recommendations are discussed.*

Keywords: Curricular programs, single-case designs, reading interventions, standard

Educators are increasingly required to implement evidence-based practices in our schools. Current legislation such as the Every Student Succeeds Act (ESSA) of 2015 (P. L. 114-95) requires the use of evidence-based activities, strategies, and interventions (U.S. Department of Education, 2016). The term *evidence-based* is defined by the U.S. Department of Education (2016) as demonstrating statistically significant effects in the improvement of student outcomes (or other relevant outcomes); evidence is considered "strong" (with at least one well-designed and well-implemented study), "moderate" (with at least one well-designed and well-implemented quasi-experimental study), or "promising" (with at least one well-designed and well-implemented correlational study controlling for selection bias). Evidenced-based practices should include a well-articulated rationale (i.e., logic model) as well.

The emphasis on the use of evidence-based practices is important to ensure efficient use of available resources and to maximize student outcomes. According to Horner et al. (2005), "Appropriate concern exists that investment in practices that lack adequate empirical support may drain limited educational resources and, in some cases, may result in the use of practices that are not in the best interests of children" (p. 175). Further, these authors noted "to support the investment in evidence-based practices, it is appropriate for any research method to define objective criteria that local, state or federal decision makers may use to determine if a practice is evidence based" (p. 175). Unfortunately, educational researchers have not discussed what research methods (i.e., group experimental, single-case) are more appropriate for validating educational practices (e.g., increasing opportunities to respond) versus curricular programs. Curricular programs include a scope and sequence of instructional skills and strategies taught through a series of lessons/exercises/tasks, typically with published materials.

Further discussion surrounding the appropriateness of particular research methods for scientifically validating curricular programs is warranted given their importance in our schools. Consider the following from the National Research Council (2004):

> Curricula play a vital role in educational practice. They provide a crucial link between standards and accountability measures. They shape and are shaped by the professionals who teach with them. Typically, they also determine the content of the subjects being taught. Furthermore, because decisions about curricula are typically made at the local level in the United States, a wide variety of curricula are available for any given subject area. Clearly, knowing how effective a particular curriculum is, and for whom and under what conditions it is effective, represents a valuable and irreplaceable source of information to decision makers, whether they are classroom teachers, parents, district curriculum specialists, school boards, state adoption boards, curriculum writers and evaluators, or national policy makers. Evaluation studies can provide that information but only if those evaluations meet standards of quality. (p. 1)

Our assumption is group experimental and single-case research methods may not be equally appropriate for validation of curricular programs.

Essentially, researchers have used two types of research methods for scientifically validating curricular programs in a quantitative manner—group experimental and single-case designs. Briefly, group experimental designs (e.g., randomized control trials [RCTs]) involve a comparison of two more equivalent groups based on differences in the independent variable. The differences *between* the groups are compared to the differences *within* the groups statistically. If the differences between the groups are greater than would be expected based on the differences within the groups, we may conclude the obtained differences were unlikely due to chance alone. The obtained mean of each group represents the average of scores and attributes of the participants in the group and, thus, allows for the evaluation of the effects of a curricular program for all group members as a whole (Lane & Gast, 2014). Two primary advantages of using group experimental designs are that cause-and-effect relationships can be determined, and it may be possible to generalize to the target population if the participants are a representative sample (Martella, Nelson, Morgan, & Marchand-Martella, 2013). However, there are also distinct disadvantages of group experimental designs. One disadvantage is the mean of a group is not representative of any one member of the group. Also, assessments are typically provided before and after the implementation of the independent variable making any adjustments to an ineffective intervention all but impossible (in other words, one may not know if the intervention is ineffective until the end of the study) (Martella et al., 2013).

Like group experimental designs, single-case designs are used to establish functional relationships. Strengths of single-case designs include (a) being useful in fields where a particular student is of concern, (b) allowing interventions to be actively monitored, and (c) being helpful in specific settings such as schools (Horner et al., 2005). Single-case designs have advantages over group experimental designs in fields focused on individuals. First, individual variability can be determined and assessed because individual participant performance is determined on a session-by-session basis via on-going data collection (Lane & Gast, 2014). Second, on-going adjustments can be made based on the obtained data if participant behavior is not changing in the desired direction (Byiers, Reichle, & Symons, 2012). Finally, there is no need for large numbers of participants to determine functional relationships (Martella et al., 2013).

Single-case studies do have a number of suggested weaknesses. For example, the ability to generalize findings beyond the few participants in the study may be limited (Maggin & Chafouleas, 2013), although single-case researchers point out external validity concerns are handled through replications (Martella et al., 2013). Another perceived weakness is that non-directly observable behaviors are considered inappropriate for applied behavior analysis (ABA) research (Critchfield & Reed, 2017). Finally, the single-case data collection must be frequent and ongoing which does not allow for larger-scale, standardized assessments to be used.

The predominant method of validating curricular programs has been group experimental designs such as RCTs (considered the "gold standard" by What Works Clearinghouse [WWC]; see Ginsburg & Smith, 2016). Although single-case designs are less utilized in the validation of such curricular programs, single-case researchers do advocate their use. For example, Horner et al. (2005) stated the following: "We provide here a context for using single-subject research to document evidence-based practices in special education…A practice refers to a curriculum, behavioral intervention, systems change, or educational approach…" (p. 175). Unfortunately,

single-case researchers have not appeared to critically analyze the use of single-case designs in curriculum validation efforts.

An obfuscating factor in the validation of curriculum programs is the lack of guidelines for such validation. Although it is obvious we should validate curricular programs, it is less obvious what is actually needed to attain this validation. We did find a study of the usefulness of RCTs in the What Works Clearinghouse (WWC) by Ginsburg and Smith (2016). They noted the following are needed when validating curricular materials when RCTs are used: (a) strong theory of why the curriculum works; (b) study is done independent of association with curriculum developer; (c) curriculum is implemented as designed; (d) comparison is identified; (e) unbiased sample has appropriate grade coverage; (f) outcomes are objectively measured, correctly analyzed, and full reported; (g) curriculum is not out of date; and (h) there is replication. The work of Ginsburg and Smith extended the report by the National Research Council (2004) on how to establish curricular effectiveness.

Similar criteria or recommendations for the use of single-case designs to validate curricular programs could not be located. While there have been several review studies assessing the effects of reading programs with various student populations with single-case designs, no studies were located that considered curricula validation issues. Studies that did consider the effects of curricular programs used accepted standards for determining the adequacy of the single-case design but not standards for validating curricular programs. For example, McKenna, Kim, Shin, and Pfannenstiel (2017) stated that research syntheses of reading practices for students with emotional and behavioral disorders (EBD) did not employ rigorous standards for single-case designs such as those outlined by the WWC: (a) independent variable is systematically manipulated, (b) dependent variable is measured by more than one assessor, (c) interobserver agreement is collected during at least 20% of data points across conditions, (d) interobserver agreement meets minimum thresholds (i.e., 80% or kappa of 6), (e) sufficient number of phases (conditions) to demonstrate an intervention effect based on design (at three different points in time or during three different phase repetitions) exist, and (f) sufficient number of data points per condition or phase (i.e., three or more) is provided. Thus, McKenna et al. (2017) reviewed reading studies using these criteria. Plavnick, Marchand-Martella, Martella, Thompson, and Wood (2015) conducted a similar investigation on reading programs with students with autism. Instead of the WWC criteria, these researchers used the criteria established by Horner et al. (2005). Both the McKenna et al. (2017) and Plavnick et al. (2015) studies came to a similar conclusion regarding the review of reading programs or practices—they hold promise for their respective populations of students. However, neither investigation considered criteria needed to validate curricular programs.

Therefore, based on what we know about curriculum development and validation coupled with the aforementioned recommendations by Ginsburg and Smith (2016), we developed a checklist of standards that are needed to establish if a curriculum has been validated above and beyond the type of research design used. (Note: the following items were not included from Ginsburg and Smith [2016]: [item a] a "strong theory of why the curriculum works" given its lack of alignment with a behavior-analytic perspective; [item d] "comparison is identified" which is not required in a single-case design; [item e] "unbiased sample" given the target population is not sampled in single-case studies although we did include appropriate grade coverage for the

amount of the program completed; [item g] "curriculum is not out of date" since this is an external validity issue; and [item h] "there is replication" as we are conducting the review on a study-by-study basis).

We did include criteria or an extension/modification of the six criteria from Ginsburg and Smith (2016). First, the authors of the study were independent of the curriculum developer(s) (item b) and/or conflict of interest procedures were followed/noted by the study authors to ensure impartiality. Second, the characteristics of the target population for which the curriculum program is designed should be specified (item c.1). The inclusion and exclusion criteria used to select the participants for evaluation studies of a curricular program should align with the characteristics of the target population for which the program was designed. Third, the level of professional development required to implement the curricular program should be documented (item c.2). Many curricular programs have focused training needs (including coaching) that should be adhered to so teachers and other curricular implementers have the level of skills needed to implement the curriculum with fidelity. At a minimum, the level of professional development received by teachers to implement the program should be documented.

Fourth, it seems clear specific lesson implementation information should be provided (cf: item c.3). For example, the duration of each lesson should be documented and compared to the lesson duration specified in the curricular program. Similarly, the activities completed in each lesson should be documented. All activities in a lesson as designed in the curricular program should be present during the investigation such as following the script (if there is one), the form of error corrections provided, and any reteaching/remediation procedures used. Horner et al. (2005) noted the need for fidelity of the intervention; however, fidelity as used by Horner et al. and others typically refers to how an intervention is provided, which does not include whether or not the intervention was applied as designed by curriculum developers. Documentation that the curricular program was implemented *as designed* is critical in the demonstration of the effects of such programs. Large deviations from the stated lesson duration are problematic. If only a few targeted activities are provided in a lesson, it is not possible to conclude what the effects of the curricular program are.

Fifth, there should be guidelines on the proportion of lessons needed to show the program works (item e). For example, is 10% of the total lessons adequate? Should we set the standard at 25%? Or, should we set the standard based on the number and range of skills covered in the program? Given that many skills covered in a program are folded into other more complex skills, it seems as if ample time should be provided to allow these skills to be developed. We propose, at a minimum, the entire program be completed if it is designed to be done within an academic year. If it is a multi-year program, we still propose assessment at the end of each academic year with a further assessment of the cumulative effects of the program over grades.

Finally, multiple measures must be included covering the range of skills taught in the program (item f). This requirement is essential given the complexity of skills needed to meet grade-level standards. It is unclear to us how multiple measures, or at least global measures representing a multitude of skills (e.g., comprehension), can be measured in a frequent enough manner in a single-case study. Further complicating this issue is the timing of the measure. In reading programs, several skills are taught at the same time, skills are folded into other more

complex skills, and many skills serve as prerequisite skills for future skills as previously described.

The purpose of this investigation was to replicate the McKenna et al. (2017) investigation using a different set of standards outlined above that we believe are critical to the validation of curricular programs. An analysis of the same studies reviewed by McKenna et al. assessing curricular programs may serve as an initial test of a proposed new set of standards we believe are needed for an effort to further validation standards.

Method

Thirty studies reviewed by McKenna et al. (2017) were analyzed based on our proposed standards for validating curricular programs. Studies involving an assessment of the effects of a curricular program were included. Of the 30 studies, 14 assessed the effects of a curricular program.

The following items were used to analyze these studies: (a) study author(s) was independent of curriculum developer(s) and/or conflict of interest procedures were followed/ noted (item b); (b) study participant(s) were consistent with the target population for which the curricular program was designed (if not, a justification for inclusion of the participant[s] was provided) (item c.1); (c) level of professional development required to implement the program was described and consistent with publisher/author guidelines (item c.2); (d) complete lessons were implemented as specified in the curricular program (item c.3); (e) 1 year of the curricular program (or complete program if less than an academic year) was completed or a multi-year program was assessed over a period of 1 year with a further assessment of the cumulative effects of the program over grades/academic years (item e); and (f) multiple measures were included that covered the range of skills taught in the program (item f).

Results

Approximately 18 programs were included in the 14 investigations. However, this finding may or may not be accurate given that the level of *PALS* used was not specified in three investigations (i.e., Barton-Arwood et al., 2005; Sutherland & Snyder, 2007; Wehby, Falk, Barton-Arwood, Lane, & Cooley, 2003) and the level (*A*, *B1*, *B2*, and *C*) and strand (*Decoding* or *Comprehension*) of *Corrective Reading* was not specified in one study (Lingo et al., 2006). The most researched program was some form of *PALS*; however, of the five investigations including *PALS*, the program was used by itself in three of these (i.e., Falk & Wehby, 2001; Lane, O'Shaughnessy, Lambros, Gresham, & Beebe-Frankenberger, 2007; Sutherland & Snyder, 2007). The most frequent measures were oral reading and nonsense word fluency. All investigations used a multiple-baseline design (MBD) in some form. Two investigations used an MBD across students, six investigations used an MBD across student pairs, and six investigations used an MBD across groups of 3-5 students. All investigations except for one (i.e., Lingo et al., 2006) used weekly probes. The authors in all investigations indicated the interventions had positive effects.

However, these conclusions may not be warranted. Using the design standards from McKenna et al. (2017), only one investigation (i.e., Barton-Arwood et al., 2005) fully met the

standards; one investigation (i.e., Cullen, Alber-Morgan, Schnell, & Wheaton, 2014) met the standards with reservations on the Ohio Achievement Assessment (OAA) and met the standards for Maze (a cloze procedure for comprehension assessment); one study (i.e., Strong, Wehby, Falk, & Lane, 2014) met the standards for fluency but not for comprehension; and 11 investigations did not meet the standards. Additionally, of the three studies that at least partially met the design standards, two did not meet the standards for overall evidence and one (i.e., Cullen et al., 2014) moderately met the standards for OAA but did not meet the standards for Maze.

Using the standards proposed in this paper, we found the following results for the 14 reviewed studies (see Table 1). None of the authors in the investigations had conflicts of interest with regard to program authorship. All but one investigation (i.e., Sutherland & Snyder, 2007) provided information justifying the inclusion of the participants in a reading program (e.g., level of reading performance). Professional development was adequately described in five investigations, partially described in one investigation, and not described in five investigations. In three investigations, professional development was described for one program but not another (when multiple programs were used).

Table 1. ***Curriculum Validation Standards Correlated to Ginsburg and Smith (2016).***

Study	item b	item c.1	item c.2	item c.3	item e	item f	Met/ Not Met Proposed Curricular Program Validation Standards	Met/ Not Met WWC Standards as assessed by McKenna et al. (2017)
Barton-Arwood et al. (2005).	Yes	Yes	Yes	No Removed seat work for *Horizons* and *PALS* modified to expand adult's role in modeling the skill and supervision (Fidelity data taken); combined both programs	No 4 days per week of *Horizons* and 3 days per week of *PALS* Note: *Horizons Fast Track AB* has 150 lessons *estimated 10-17 weeks of intervention (number of lessons completed unspecified)	Yes Note: Benchmark not assessed; standardized pre-test/posttest scores reported	No	*Design Standards*: Yes *Overall Evidence:* No for all measures
Cullen et al. (2014).	Yes	Yes	No	Yes (fidelity data taken)	No Completed 8-15 of 50 lessons *estimated 8-15 sessions	Yes Note: Benchmark not assessed	No	*Design Standards*: With Reservations for OAA Yes for Maze *Overall:* Moderate for OAA No for Maze
Falk & Wehby (2001).	Yes	Yes	No	Yes (fidelity data taken)	No *K-PALS* total for 11 weeks *estimated 3-10 weeks of peer tutoring component, 9-30 lessons (number of lessons completed unspecified)	Yes Note: Benchmark not assessed	No	*Design Standards*: No *Overall Evidence:* N/A

Study	item b	item c.1	item c.2	item c.3	item e	item f	Met/ Not Met Proposed Curricular Program Validation Standards	Met/ Not Met WWC Standards as assessed by McKenna et al. (2017)
Harris, Oakes, Lane, & Rutherford (2009).	Yes	Yes	Yes for Sonday No for Great Leaps	Yes for *Sonday Reading Program* (fidelity data taken); No on *Great Leaps*	No 27 to 47 sessions of instruction; took 2 to 3 sessions to complete a lesson (number of lessons completed unspecified)	Yes Note: Results compared to benchmark level	No	*Design Standards*: No *Overall Evidence:* N/A
Lane, Little, Redding-Rhodes, Phillips, & Welsh (2007).	Yes	Yes	Yes	Yes (fidelity data taken)	No 9 weeks of lessons (number of lessons completed unspecified)	No, did not assess all skills taught in program such as comprehension Note: Benchmark not assessed; reported slopes	No	*Design Standards*: No *Overall Evidence:* N/A
Lane et al. (2001).	Yes	Yes	Yes	Yes (fidelity data taken)	No 10 weeks, 30 lessons; *PATR* program takes 12-14 weeks 3-4 times per week	No, did not directly assess phonological awareness Note: Normative/ ambitious growth levels compared	No	*Design Standards*: No *Overall Evidence:* N/A

Table Continues

Study	item b	item c.1	item c.2	item c.3	item e	item f	Met/ Not Met Proposed Curricular Program Validation Standards	Met/ Not Met WWC Standards as assessed by McKenna et al. (2017)
Lane et al. (2002).	Yes	Yes	No	Yes (fidelity data taken)	No 9 weeks, 30 lessons (number of books completed out of six unspecified)	No, measures did not include all program components such as spelling Note: Benchmark not assessed; calculated individual effect sizes	No	*Design Standards*: No *Overall Evidence:* N/A
Lingo, Slaton, & Jolivette (2006).	Yes	Yes	No	Yes (fidelity data taken but not specified)	No Completed 5-19 lessons	Yes Note: Benchmark not assessed; reported standard scores and grade equivalents	No	*Design Standards*: No for both measures *Overall Evidence:* N/A
Oakes, Mathur, & Lane (2010).	Yes	Yes	Yes for *Foundations*, No for *Harcourt Trophies* and *Voyager's Blastoff to Reading* Program	Yes (fidelity data taken)	No Primary program: *Harcourt Trophies* for 6 weeks Secondary program (baseline): *Foundations* for 6-10 weeks) Secondary program with *Voyager's Blastoff* (experimental): 8 weeks) (number of lessons unspecified)	Yes, only for *Voyager's Blastoff* Note: Reported slopes and realistic and obtained growth compared to realistic and ambitious gains	No	*Design Standards*: No *Overall Evidence:* N/A

Table continues

Study	item b	item c.1	item c.2	item c.3	item e	item f	Met/ Not Met Proposed Curricular Program Validation Standards	Met/ Not Met WWC Standards as assessed by McKenna et al. (2017)
Scott & Shearer-Lingo (2002).	Yes	Yes	No	Yes (no fidelity data reported)	No *Teach Your Child to Read in 100 Easy Lessons* implemented for approx. 2 weeks *estimated: *Great Leaps* in effect up to 35 days (number of lessons completed unspecified)	No, only fluency measures Note: Benchmark not assessed	No	*Design Standards*: No *Overall Evidence:* N/A
Strong et al. (2004).	Yes	Yes	Yes	Yes (fidelity data taken)	No 7 –weeks of training, *estimated 28 lessons of *Corrective Reading* (number of lessons completed unspecified)	Yes Note: Benchmark not assessed; standardized pre-test/posttest scores reported	No	*Design Standards*: Yes for Fluency, No for comprehension *Overall Evidence:* No for both measures
Sutherland & Snyder (2007).	Yes	No	No—Unspecified type and length of training	Yes for *PALS* and self graphing added (fidelity data taken)	No *estimated 2 to 6 weeks of implementation (number of lessons completed unspecified)	No, only fluency Note: Reported slopes and compared obtained fluency scores to a goal increase of 1.39 wpm per week.	No	*Design Standards*: No *Overall Evidence:* N/A

Table Continues

Study	item b	item c.1	item c.2	item c.3	item e	item f	Met/ Not Met Proposed Curricular Program Validation Standards	Met/ Not Met WWC Standards as assessed by McKenna et al. (2017)
Wehby et al. (2003).	Yes	Yes	Yes	No (fidelity data taken)	No *estimated 6-9 weeks or 24-36 lessons (number of lessons completed unspecified)	No, only fluency Note: Benchmark not assessed; standardized pre-test/posttest scores reported	No	*Design Standards*: No *Overall Evidence:* N/A
Wehby, Lane, & Falk (2005).	Yes	Yes	No for *Scott Foresman* Yes for *PATR*	Yes (no fidelity data for *Scott Foresman*; fidelity data taken for *PATR*)	No *estimated 3-8 weeks (4 days per week) of *Scott Foresman* alone (baseline); 9 weeks (3-4 days per weeks, 32 lessons) of *Scott Foresman* and *PATR* combined (intervention)	No, only fluency Note: Benchmark not assessed; reported slopes	No	*Design Standards*: No *Overall Evidence:* N/A

Note: item b = Study independence/ conflict of interest procedures were followed/noted; item c.1 = Justification for inclusion of participant(s) provided; item c.2 = Professional development described; item c.3 = Complete lessons implemented as specified; item e = 1 year of academic program completed/multi-year program assessed over 1 year with assessment of cumulative effects over grades; item f = Multiple measures included covered range of skills taught

Information on how complete lessons were implemented as specified was found in 12 investigations with fidelity data taken in all but one (i.e., Scott & Shearer-Lingo, 2002). Of the 11 investigations including fidelity data, partial fidelity data (one program implementation was evaluated but not the another in a multiple program implementation) were taken in two investigations (i.e., Harris et al., 2009; Wehby et al., 2005) and one investigation (i.e., Lingo et al., 2006) had fidelity data but the data were not reported. Programs were not implemented as specified but fidelity data were taken in two investigations (i.e., Barton-Arwood et al., 2005; Wehby et al., 2003).

Programs were not implemented for at least a 1-year period nor were any programs completed in their entirety as far as we could determine. Programs were implemented over several weeks (we estimated the longest implementation was up to 17 weeks—Barton-Arwood et al., 2005). The number of lessons completed were as few as 5 (Lingo et al., 2006) to as many as 32 (Wehby et al., 2005). However, these numbers may not be accurate given that when the

authors reported the number of lessons they may have meant the number of instructional sessions; lessons may take more than one session to complete.

Complicating matters further, the number of lessons completed from a program was not specified in 10 investigations. Only Cullen et al. (2014) (8-15 of 50 *Headsprout* lessons completed), Lane et al. (2001) (30 of an estimated 40 plus lessons of *Phonological Awareness Training for Reading* [*PATR*] completed), Lingo et al. (2006) (completed 5-19 lessons of 60 to 140 lessons of *Corrective Reading*—depending on the level and strand of the program), and Wehby et al. (2005) (completed 32 lessons of combined programs that had 40 plus lessons for *PATR* to full year/multi-grade lessons for *Scott Foresman*) explicitly stated the number of program lessons completed.

Seven investigations did not include measures assessing skills taught—typically these investigations included fluency measures such as oral reading and nonsense word fluency even though other skills were taught in the program(s) such as spelling and comprehension. Six of the investigations included multiple measures covering the skills taught, and one investigation (i.e., Oakes et al., 2010) included measures for only the *Voyager Blastoff to Reading* program but not for the other programs. However, there is a caveat here. Even though these seven investigations used measures that assessed skills taught, benchmark analyses were not conducted. Benchmark assessments provide a minimum threshold for grade level performance. Only four investigations reported or mentioned benchmark or expected fluency growth per week (i.e., Harris et al., 2009; Lane et al., 2001; Oakes et al., 2010; Sutherland & Snyder, 2007). Thus, although technically, the investigation may get a "Yes" in this category, progress monitoring of skills taught in a program does not necessarily indicate if educationally and socially significant progress has been achieved. Of the 10 investigations that did not report or compare results to benchmarks in reading, three reported standard score data, two reported slope data, one reported effect sizes, and one reported standard scores and grade equivalents. Based on the standards proposed in this paper, we concluded none of the investigations met standards required to validate a curriculum program in reading.

Discussion

Our analysis of 14 of the 30 articles analyzed by McKenna et al. (2017) revealed none of the investigations were able to establish a curriculum as effective based on our proposed standards. Interestingly, none of the investigations we reviewed provided evidence of effectiveness in the McKenna et al. review either. McKenna et al. found only two reading interventions were found to be potentially promising—cognitive mapping and listening while reading—neither of which are programs and were not reviewed here. According to McKenna et al., "findings from this review suggest there continues to be a lack of evidence-based reading practices for students with and at risk for EBD, limiting the ability of research to inform professional development and training" (p. 898). This issue is not only true of investigations including students with emotional and behavioral disorders. Perhaps the reason for a lack of such research was the methodology was not suited for such an endeavor.

There was a general theme we found in the 14 articles included in this paper, some of which were due to experimenter error/oversight and some due to the constraints of single-case designs. First, there was a general failure to complete a full program or at least to provide

detailed information on program fidelity related to number of lessons completed, sessions needed to complete lessons, and number of instructional sessions per day. It is simply not possible to replicate most of these investigations given the lack of information provided.

Second, the number of lessons included in a program was rarely stated in an investigation. There is simply no way to evaluate study results and conclusions regarding the effectiveness of a program without knowing the extent of the program completed. Similarly, specifics of a program such as level (e.g., *Corrective Reading Decoding B1*) or strand (e.g., *Decoding* or *Comprehension*) were not always stated. To indicate *Corrective Reading* was implemented without specifying which level and strand was used prevents us from making any conclusions on program effectiveness.

Third, many investigations included the use of multiple programs implemented at the same time. The only conclusion about effectiveness that can be made is with regard to the combined effects of the programs; no conclusions can be made regarding individual program effects. This problem is also seen when programs and/or additional instructional time are added. For example, Strong et al. (2004) began with 25 min of instruction in baseline (10 min of writing in journals and 15 min of taking turns reading a story aloud; note: other activities such as spelling were provided but the amount of time was not specified) then implemented 30-40 min of *Corrective Reading Decoding Level B1*. Following this, *Great Leaps Reading Stories* was implemented adding another 20-30 min of instruction. Therefore, it is not possible to conclude either program had an effect given the amount of reading instruction greatly increased as well.

Fourth, given that single-case designs require frequent repeated measures, it is not surprising reading fluency measures were used in all but one investigation (i.e., Cullen et al., 2014). However, many of the programs found in these investigations taught more than just reading fluency. Measures of reading comprehension often were not used. In fact, reading comprehension was measured in only three investigations (i.e., Barton-Arwood et al., 2005; Cullen et al., 2014; Strong et al., 2004). There was also a lack of reporting of benchmark scores—scores that are valuable to educators and show how instruction closes the gap in reading skills achieved by grade-level peers.

Based on our analysis and experience with the development and validation of curricular materials, we have come to the conclusion that single-case research methodology by itself may not be adequate to validate curricular programs, or at least has not been shown to be adequate at this point. We, as behaviorally-oriented researchers, should look outside our own methodology to answer and address issues we have neglected for far too long. We agree with Horner et al. (2005) that:

> The selection of any research methodology should be guided, in part, by the research question(s) under consideration. No research approach is appropriate for all research questions, and it is important to clarify the types of research questions that any research method is organized to address. (p. 172)

In his treatment of the recommendations by Bear, Wolf, and Risley (*BWR*) in 1968, Axelrod (2017) stated the following: “If one of the research designs recommended by [*BWR*] is not

feasible, and a group–comparison design is possible, researchers should use it without hesitation" (p. 169).

Perhaps because of the adherence by behavior analysts to the Essential Characteristics of ABA as outlined by *BWR*, there is a noticeable lack of research in the technology of teaching, specifically, curriculum development. As argued by Critchfield and Reed (2017), these seven characteristics (i.e., applied, behavioral, analytical, conceptually systematic, effective, and generality) may have served to hold back ABA research in important areas. The framework developed by *BWR* "fueled the growth of ABA. Ironically, however, in contemporary use, the framework serves as a bottleneck that prevents many socially important problems from receiving adequate attention in applied behavior analysis research" (p. 123) or require researchers to use research methods ill fitted to the question to be answered. We believe development and validation of curricular programs is one such example. Unfortunately, it seems as if cognitive scientists have the corner on the learning sciences, which is ironic given that Skinner described an explicit and systematic method of teaching described in his book, *The Technology of Teaching* (1968). Those in ABA appear to have turned over this technology to cognitive scientists who are generally seen as the "experts" in instruction, rather than modifying how they conduct their own research. Critchfield and Reed (2017) listed several areas of research adversely affected by strict adherence to the seven characteristics including, for example, (a) research on voucher systems to reduce workplace attendance and drug usage by drug abusers and (b) Positive Behavior Interventions and Supports. We would add curriculum validation to this list and advocate behavior analysts use research methods that can determine the effects of curriculum programs, if done so appropriately. We believe Axelrod (2017) sums up our position when referring to adherence of the seven ABA characteristics.

> My recommendation to ABA researchers on this issue is *not* to consider the measurement procedure or the research design as the most critical parts of a study; instead, they should regard the research *question* as the most important aspect of any study. Next, researchers should use the most scientific measurement procedures and the best research designs that are feasible (p. 168).

This recommendation is the same one we advocated for in our research methods textbook (Martella et al., 2013). We fear that a reason why ABA researchers have fallen behind other researchers in the learning sciences is because they do not consider research questions and/or do not use adequate research designs to answer them. In conclusion, we agree with Axelrod's (2017) answer to whether ABA researchers should abandon questions not in strict adherence to the requirements set by *BWR* is "Absolutely not!... If one of the research designs recommended by *BWR* is not feasible, and a group–comparison design is possible, researchers should use it without hesitation" (p. 169).

References

References marked with an asterisk (*) indicate studies included in the synthesis by McKenna et al. (2017) and analyzed in this study.

Axelrod, S. (2017). A commentary on Critchfield and Reed: The fuzzy concept of applied behavior analysis research. *The Behavior Analyst, 40*, 167-171. doi:10.1007/s40614-017-0117-6

*Barton-Arwood, S., Wehby, J., & Falk, K. (2005). Reading instruction for elementary-age students with emotional and behavioral disorders: Academic and behavioral outcomes. *Exceptional Children, 72*, 7-27.

Bear, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97.

Byiers, B. J., Reichle, J., & & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*(4), 397-414. doi:10.1044/1058-0360(2012/11-0036)

Critchfield, T. S., & Reed, D. D. (2017). The fuzzy concept of applied behavior analysis research. *The Behavior Analyst, 40,* 1-37. doi:10.1007/s40614-017-0093-x

*Cullen, J., Alber-Morgan, S., Schnell, S., & Wheaton, J. (2014). Improving reading skills of students with disabilities using Headsprout Comprehension. *Remedial and Special Education, 35*, 356-365. doi:10.1177/0741932514534075

*Falk, K., & Wehby, J. (2001). The effects of peer-assisted learning strategies on the beginning reading skills of young children with emotional or behavioral disorders. *Behavioral Disorders, 26*, 344-359.

Ginsburg, A., & Smith, M. S. (2016). *Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the What Works Clearinghouse.* Washington, DC: American Enterprise Institute. Retrieved from https://www.carnegiefoundation.org/resources/publications/do-randomized-controlled-trials-meet-the-gold-standard/

*Harris, P., Oakes, W., Lane, K., & Rutherford, R. (2009). Improving the early literacy skills of students at-risk for internalizing or externalizing behaviors with limited reading skills. *Behavioral Disorders, 34*, 72-90.

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*(2), 165-179. doi:10.1177/001440290507100203

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*, 445-463. doi:10.1080/09602011.2013.815636

*Lane, K., Little, M., Redding-Rhodes, J., Phillips, A., & Welsh, M. (2007). Outcomes of a teacher-led reading intervention for elementary students at-risk for behavioral disorders. *Exceptional Children, 74*, 47-70. doi:10.1177/001440290707400103

*Lane, K., O'Shaughnessy, T., Lambros, K., Gresham, F., & Beebe-Frankenberger, M. (2001). The efficacy of phonological awareness training with first-grade students who have behavior problems and reading difficulties. *Journal of Emotional and Behavioral Disorders, 9*, 219-231. doi:10.1177/106342660100900402

*Lane, K., Wehby, J., Menzies, H., Gregg, R., Doukas, G., & Munton, S. (2002). Early literacy instruction for first-grade students at-risk for antisocial behavior. *Education and Treatment of Children, 25*, 438-458.

*Lingo, A., Slaton, D., & Jolivette, K. (2006). Effects of Corrective Reading on the reading abilities and classroom behaviors of middle school students with reading deficits and challenging behavior. *Behavioral Disorders, 31*, 265-283. doi:10.1177/019874290603100305

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the special series: Issues and advances of synthesizing single-case research. *Remedial and Special Education, 34*(1), 3-8. doi:10.1177/0741932512466269

Martella, R. C., Nelson, J. R., Morgan, R. L., & Marchand-Martella, N. E. (2013). *Understanding and interpreting educational research.* New York, NY: Guilford Press.

McKenna, J. W., Kim, M. K., Shin, M., & Pfannenstiel, K. (2017). An evaluation of single-case reading intervention study quality for students with and at risk for emotional and behavioral disorders. *Behavior Modification, 41*(6), 868-906. doi:10.1177/0145445517701896

National Research Council. (2004). *On evaluation curricular effectiveness: Judging the quality of K-12 mathematics evaluations.* Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials. Mathematical Sciences Education Board, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

*Oakes, W., Mathur, S., & Lane, K. (2010). Reading interventions for students with challenging behavior: A focus on fluency. *Behavioral Disorders, 35*, 120-139.

Plavnick, J. B., Marchand-Martella, N. E., Martella, R. C., Thompson, J. L., & Wood, A. L. (2015). A review of explicit and systematic scripted instructional programs for students with autism spectrum disorder. *Review Journal of Autism and Developmental Disorders, 2*(1), 55-66. doi:10.1007/s40489-014-0036-3

*Scott, T., & Shearer-Lingo, A. (2002). The effects of reading fluency instruction on the academic and behavioral success of middle school students in a self-contained EBD classroom. *Preventing School Failure, 46*, 167-173. doi:10.1080/10459880209604417

Skinner, B. F. (1968). *The technology of teaching.* New York, NY: Appleton-Century-Crofts.

*Strong, A., Wehby, J., Falk, K., & Lane, K. (2004). The impact of a structured reading curriculum and repeated reading on the performance of junior high students with emotional and behavioral disorders. *School Psychology Review, 33*, 561-581.

*Sutherland, K., & Snyder, A. (2007). Effects of reciprocal peer tutoring and self-graphing on reading fluency and classroom behavior of middle school students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 15*, 103-118. doi:10.1177/10634266070150020101

U.S. Department of Education. (2016). *Nonregulatory guidance: Using evidence to strengthen education investments.* Washington, DC: Retrieved from *https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf*

*Wehby, J., Falk, K., Barton-Arwood, S., Lane, K., & Cooley, C. (2003). The impact of comprehensive reading instruction on the academic and social behavior of students with emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 11*, 225-238. doi:10.1177/10634266030110040401

*Wehby, J., Lane, K., & Falk, K. (2005). An inclusive approach to improving early literacy skills of students with emotional and behavioral disorders. *Behavioral Disorders, 30*, 155-16.