Hammill Institute on Disabilities

Behavioral Disorders
2017, Vol. 42(2) 52–64

DOI: 10.1177/0198742916688653
journals.sagepub.com/home/bhd
SAGE

Quality Indicators for Systematic Reviews in Behavioral Disorders

Daniel M. Maggin, PhD[1], Elizabeth Talbott, PhD[1], Eryn Y. Van Acker, MA[1], and Skip Kumm, MA[1]

Abstract

Special education researchers, practitioners, and policy makers continue to work toward developing and implementing empirically supported practices and policies to address the academic, social, and postschool challenges confronting students with emotional and behavioral disorders. The systematic review has become an essential vehicle for compiling and disseminating research findings on an array of topics. Given the importance and impact of these research summaries, it is instructive to take stock of the extent to which reviews in our field adhere to current standards. Drawing from a number of sources in the social and behavioral sciences, we propose and describe a series of quality indicators for systematic reviews. We applied these indicators to systematic reviews published in *Behavioral Disorders* between 2005 and 2016 with the broader goal of identifying areas of methodological strength and areas for improvement. Results indicate that the sample of systematic reviews demonstrated particular strength in several procedural domains such as the specification of inclusion criteria, identifying the electronic databases used for the search, and describing the plan for data analysis. We also identified a number of areas to which researchers might devote greater attention to increase the rigor of systematic reviews in the field. Findings are contextualized within the importance of research transparency and reporting to improve practice and policy.

Keywords

emotional and behavioral disorders, quality indicators, meta-analysis, systematic review

The development, identification, and dissemination of evidence-based intervention and assessment strategies have been a dominant theme within education for more than a decade (Shavelson & Towne, 2002). Following the logic of evidence-based practice, rigorous research methods are used to determine which strategies are most likely to produce the desired outcomes (Gersten et al., 2005). A critical aspect of the identification of evidence-based educational methods is to reliably summarize the research on a given topic to assess the methodological strength and consistency of findings to draw meaningful conclusions. As such, the systematic review has gained increased attention for its emphasis on providing transparent assessments of research (Cook, Tankersley, & Landrum, 2009). Given the variety of academic, social, and postschool challenges confronting students with emotional and behavioral disorders (EBD), systematic research reviews can have significant implications and reach for guiding the adoption and implementation of various educational practices and policies for this particularly vulnerable population of students. For instance, the results of systematic reviews can be used by practitioners to select intervention methods, policy makers to weigh the costs and benefits of adopting particular strategies, researchers to identify the strengths and weaknesses within a body of research to assist with planning their next study, grant funders to determine issues that are in need of additional attention and resources, and the developers of standards to detect promising strategies and practices (Talbott, Maggin, Van Acker, & Kumm, in press). The strength of systematic reviews derives from the articulation of a clear, transparent, replicable methodology that increases the likelihood of drawing objective conclusions from the research being reviewed. As such, the systematic review has become an essential tool within the development and dissemination of empirically supported practices and policies for students with EBD (Maggin & Chafouleas, 2013).

Despite the importance of systematic reviews for understanding and addressing an array of complex issues for students with EBD, field-specific guidance for researchers or practitioners on this research methodology is currently unavailable. The purpose of this article, therefore, is to follow the lead of previous efforts in the field to establish guidelines and provide an overview of common quality indicators for conducting systematic reviews (Odom et al., 2005). In the following sections, we define systematic reviews in relation to other types of literature reviews, describe common quality indicators for systematic reviews,

[1]The University of Illinois at Chicago, USA

Corresponding Author:
Daniel M. Maggin, The University of Illinois at Chicago, 3424 ETMSW, 1040 W Harrison Street, Chicago, IL 60607, USA.
Email: dmaggin@uic.edu

and provide a conceptual discussion of the reasons these standards are important. We then apply these indicators to systematic reviews published in *Behavioral Disorders* from 2005 to 2016 to identify areas of methodological strength and areas for improvement in recent reviews in the field. The article concludes with recommendations for improving the overall quality of systematic reviews in our discipline, to ensure that the most valid information is disseminated for students with EBD who have among the most unique and complex needs in special education (Maggin, Wehby, Farmer, & Brooks, 2016).

Defining Systematic Reviews

A systematic review is the attempt to systematically collate all empirical evidence that meets a set of specific eligibility criteria with the purpose of answering a particular research question (Lipsey & Wilson, 2001). As such, these research syntheses require the development of clearly articulated procedures that are selected to minimize subjectivity and maximize transparency, which increase the likelihood of providing reliable findings from which objective conclusions can be drawn to support decision making for policy and practice (Borenstein, Hedges, Higgins, & Rothstein, 2009). It is essential not only to define systematic reviews but also to distinguish them from other types of literature reviews used in our field. For instance, systematic reviews are fundamentally different from narrative reviews and position papers in that researchers describe and use a replicable, transparent set of procedures to search, include, and review the research on a given topic. In contrast, narrative reviews typically do not provide details on how particular decisions are made regarding the relevance and consideration of the included studies (Baumeister & Leary, 1997). This is not to discount the importance of narrative reviews; in fact, narrative reviews can often be a means for disseminating powerful scholarship and perspective. Another type of review that is common in our field, and one that is closely related to systematic reviews, is the meta-analysis. There is often confusion, however, regarding the differences between a meta-analysis and a systematic review. Meta-analysis refers explicitly to the use of statistical methods to summarize and combine the results of independent research studies (Cooper, Hedges, & Valentine, 2009). As such, a systematic review does not require the use of meta-analytic techniques although the validity of findings from a meta-analysis is directly related to the quality of the systematic review procedures used to reduce bias (Lipsey & Wilson, 2001).

Guidelines for Systematic Reviews

Part of the motivation for the current article is to provide researchers and research consumers with a set of guidelines to use when conducting and reviewing systematic reviews. These efforts are similar to the development of indicators for primary research proposed by our esteemed colleagues in special education more than a decade ago (Odom et al., 2005). The purpose of those methodological guidelines, as with those proffered here, was to encourage the use of strong research methods to ensure the most valid information is obtained and disseminated. It is important to note that the development of quality indicators has not been isolated to the execution of primary research studies in the field of special education. In fact, there have been several efforts to establish guidelines for conducting systematic reviews in other areas of the social and behavioral sciences (e.g., American Psychological Association [APA], 2010; Higgins & Green, 2011; Moher et al., 2015). These endeavors have led to the development of a set of widely accepted standards on the processes and procedures for conducting a rigorous systematic review. In the current article, we synthesize extant guidelines to generate a set of quality indicators for conducting and reporting rigorous systematic reviews in special education generally and in the field of EBD specifically.

Quality Indicators for Systematic Reviews

In the following sections, we provide an overview of the quality indicators for conducting systematic reviews. Each criterion and its operational definition are presented in Table 1.

Research Question

As with all types of research, a rigorous systematic review begins with the articulation of the research question that communicates the primary objectives of the review (Cooper et al., 2009). Because systematic reviews can address a wide array of topics and include a range of research designs, there are few formal criteria for presenting the research question itself. Recommendations on the development of research questions for systematic reviews, however, emphasize the need for them to be testable and to include explicit reference to the key populations, variables, and objectives of the research to be sampled (Cooper et al., 2009). That is, the research questions should unambiguously refer to the necessary characteristics of the populations being sampled, primary variables of interest, and intent of the review. The overarching goal of the research question is to clearly communicate the population of studies that will be eligible for the review to the readership (Cochrane Collaboration, 2013). For instance, if the purpose of the review is to determine whether a particular intervention program has sufficient empirical support to warrant its use for reducing anxiety experienced by students with EBD, the research question would reference the particular age and grade levels of interest, the particular name or class of interventions that will be eligible for the review, and the specific outcome measures considered. Similarly, if the systematic review will examine the perceptions of students and school personnel on the

Table 1. Methodological Quality Indicators for Conducting Systematic Reviews.

Methodological Domain	Quality indicator
Research question	
Research question formulation	The research question communicates the objectives and provides boundaries for making decisions about which studies to include in the systematic review.
Eligibility criteria	
Variable characteristics	Operational characteristics of the variables that will be the focus of the review are described with examples of key variables such as the interventions, assessments, programs, practices, or policies of interest.
Participant information	Key participant information is reported including any disability, demographic, and/or functional characteristics that define the student population.
Research design	The research designs eligible for the review are identified with key methods operationalized if necessary.
Time period	The time period in which the research had to be published to be eligible was identified.
Search procedures	
Databases identified	The electronic reference databases searched were referenced directly.
Registries identified	The prospective research registries searched were referenced directly.
Unpublished sources	The authors reported whether unpublished studies were included in the search.
Keywords identified	Keywords used to search reference databases and registries were explicitly identified.
Search date	The date on which the final search was conducted was reported.
Hand search conducted	The authors reported whether journals were hand searched and, if so, which journals.
Citation lists review	The authors reported whether citation lists of included studies were reviewed and, if so, which studies.
Subject matter expert review	The authors reported whether individuals with expertise on the subject matter of the review examined the citation list to assist with identifying additional, relevant studies.
Review articles consulted	The authors reported whether pertinent review articles were reviewed for citations.
Authors contacted	The authors reported whether authors were contacted to see if other reports were available.
Language specification	The authors specified the languages of eligible reports and if there were any limitations imposed on the language of the report.
Titles and abstracts reviewed	The authors reported whether titles and abstracts were used to determine eligibility in the study.
Searcher qualification	The authors reported the qualifications for those reviewing studies for eligibility.
Search agreement	The authors reported agreement across those reviewing for study eligibility.
Retrieval procedures	
Total citations return	The authors reported the number of studies identified through the initial data-based search (i.e., total returns from search).
Total screened out	The authors reported the number of studies excluded during the initial database search (i.e., how many relevant studies were there to retrieve from the initial number of citations).
Total retrieved	The authors reported the number of studies identified as potentially eligible that were successfully retrieved for full review.
Total excluded	The authors reported the number of studies that were excluded from those after the retrieval process due to ineligibility.
Systematic screening	
Reasons studies excluded	The authors reported the reasons excluded studies were deemed ineligible.
Total studies included in review	The authors reported the total number of studies that met eligibility criteria and were subsequently included in the review.
Coder training and expertise	The authors reported the training and expertise of those charged with conducting the screening process.
Reliability reported	The authors reported the reliability or interobserver agreement statistics used to evaluate the consistency of the screening process.
Disagreement resolution method	The authors reported the method used to resolve any disagreements among screeners.
Coding scheme procedures	
Coder expertise	The authors reported the expertise of individuals charged with coding studies.
Coder training	The authors reported methods for training individuals charged with coding studies.
Proportion double coded	The authors reported the number and proportion of studies that were coded by more than one independent observer.

(continued)

Table 1. (continued)

Reliability reported	The authors reported the reliability or interobserver agreement statistics used to evaluate the consistency of the coding.
Disagreements resolution method	The authors reported the procedures used to resolve disagreements among coders.
Response categories reported	The authors reported the response categories available for coders to select from.
Missing information process	The authors reported the procedures used to address issues for missing information within research studies, such as checking with authors on missing information.
Coding scheme content	
Participants characteristics	The authors collected data pertaining to participants.
Key variable features	The authors collected data pertaining to variables under study.
Methodological quality	The authors collected data pertaining to study quality.
Data analysis plan	
Data analysis plan provided	The authors provide a data analysis plan.
Coding scheme aggregation specified	The data analysis plan includes the method used to descriptively aggregate the data from the coding scheme.
Results method described	The data analysis plan includes the method used to synthesize the results reported across studies.

implementation of a particular school discipline policy for students with EBD, the review authors might indicate in the research question that qualitative research literature on this topic will be reviewed in addition to sample and variable characteristics that establish the boundaries of the studies to be included. Regardless of the overall purpose of the review, the research question should be constructed so that it serves as an advanced organizer to the reader and communicates the primary aims of the review and the key details of the studies to be included.

Eligibility Criteria

Following the development of a research question that provides readers with the critical details and goals of the review, it is necessary to develop a set of criteria that clearly defines the population of studies from which the research team will ultimately sample (Lipsey & Wilson, 2001). Whereas the research questions are meant to serve as an advanced organizer for the reader, the eligibility criteria provide the operational characteristics of the studies that will be included and excluded. The development and reporting of clear eligibility criteria not only benefits the research team in preparing a transparent and replicable review but also provides readers with the opportunity to determine whether the universe of eligible studies was successfully secured. As such, the audience is then in a position to evaluate whether the research question was adequately addressed and whether the main objectives of the review were successfully achieved. Put simply, the development and reporting of clear and replicable eligibility criteria is essential to both planning the research as well as evaluating the overall validity of the review.

The specific details of what should be included in the eligibility criteria will naturally depend on the purpose of the review. However, most systematic reviews should address information such as (a) the demographic, emotional, behavioral, and academic information that distinguishes the student population of interest; (b) the contexts and settings in which the research has taken place; (c) the pertinent constructs and outcomes and whether there are any restrictions on how these were measured; and (d) the research designs or other methodological characteristics considered when identifying studies (Davies, 2000). Providing detailed inclusion criteria for the particular sample characteristics of interest can be particularly important given the range and complexity of the emotional and behavioral challenges experienced by students with EBD (e.g., Maggin et al., 2016). As such, the more specific that review authors can be regarding the focus of the research, the more specific the subsequent conclusions and recommendations can be as well. Moreover, it is often necessary to specify methodological criteria to clearly communicate the types of research designs eligible for addressing the review questions. Review of intervention research for students with EBD in particular often requires noting the specific research designs that were eligible for review given the range of experimental frameworks used across the literature. As such, many systematic reviews will stipulate that only those studies that randomly assigned participants to intervention or control conditions, used a particular quasi-experimental framework, or applied single-case research methods would be eligible (Borenstein et al., 2009). Ultimately, the development of the specific eligibility criteria for each review will depend on its overall aims and purpose. It is important to note, however, that precise eligibility criteria can assist researchers in executing a thorough and efficient search process as well as making the review transparent, replicable, and interpretable.

Search Procedures

Once the eligibility criteria have been established, the systematic review process moves away from conceptual

development toward implementation. The purpose of the search is to identify the universe of eligible studies that meet the eligibility criteria. It is important to note that the strength of the review largely depends on the extent to which all the relevant studies are successfully identified and subsequently retrieved. Each individual study that is included in the review contributes new information either by providing additional support or contrary evidence to the results of the other studies (Borenstein et al., 2009). Because each research report contributes to the evidence base, the results and conclusions of a systematic review are related to the extent to which all eligible research is considered. Researchers must therefore clearly describe the procedures for searching the literature to ensure all relevant research reports are identified. Many common strategies for searching the literature base for studies are presented in the proposed quality indicators in Table 1. Examples of these procedures include explicitly naming the electronic databases, including databases of unpublished studies and, if applicable, research registries (i.e., publicly available database of ongoing and completed trials) searched; listing the keywords used to identify reports within the electronic searches; and describing any alternative procedures used to identify eligible reports such as hand searching journals, examining the citation lists of previous reviews, or contacting authors and other subject matter experts.

Providing a description of the procedures taken to identify studies enhances the transparency of the process for readers and represents an important aspect of systematic reviews. In addition to these procedural steps, a number of additional reporting standards address the extent to which the pool of studies is reflective of the universe of possible studies on the topic. For instance, reporting the qualifications of those individuals conducting the search is important for determining whether there was sufficient expertise and experience, indicating whether there were language barriers that precluded the identification and subsequent review of studies in languages other than English can be an important limitation in many systematic reviews, and providing the date of the final search allows consumers to determine whether the current review is up-to-date or whether additional research might have been published on the topic. Because the intentional or unintentional exclusion of relevant studies raises the possibility that the review findings are not representative of the broader population of research findings, the reporting of these steps is essential to drawing valid and transparent conclusions.

Retrieval Procedures

In addition to well-described search procedures, it is also important to provide a transparent and replicable description of the retrieval process (Higgins & Green, 2011). Researchers cannot retrieve a portion of the identified studies in many cases. The specific reasons for not obtaining copies of identified reports can vary although some common reasons are that the particular record is unavailable either online or in print, the original document was lost or harmed, or the research report is in a different language. Regardless of the reason, there is almost always a portion of research reports that cannot be acquired. As such, it is important for review authors to record those studies that were deemed either eligible or possibly eligible but were not retrievable. Cataloging the number and proportion of papers that were retrieved provides readers with vital information to evaluate the rigor and validity of the review. Put another way, the validity of the results is compromised when only the easy-to-find studies are included (Gough, Oliver, & Thomas, 2013). Reviews should contain explicit reference to the proportion of potentially eligible studies not obtained during the search with reasons for their omission.

Coding Scheme Development and Implementation

The research questions in a systematic review are often addressed—in full or in part—through the development of a coding scheme. These coding schemes are used to collect descriptive information from the studies that can then be used to characterize the samples, outcome, and other key variables of interest (Lipsey & Wilson, 2001). For example, in their review of disproportionality of English Language Learners with EBD, Gage, Gersten, Sugai, and Newman-Gonchar (2013) collected demographic information that not only described the samples but also were used as variables in the subsequent meta-analysis to examine variability in disproportionality rates. To transparently and reliably collect this information, the authors need to develop a coding protocol, which in turn guides the information they collect. It should be emphasized that the coding protocol developed by Gage et al. was developed to address their research questions pertaining to the extent of disproportionality and investigate the underlying patterns. Because coding schemes are developed to address the research questions of a given study, the particulars of each protocol will inevitably differ. Despite this inherent variability, specific coding categories can be employed in a majority of reviews, given their importance for evaluating both the rigor and generality of research drawn from various paradigms. These general coding categories and their content are reviewed in the following sections.

Methodological quality. Methodological quality refers broadly to the degree to which the research design procedures generate valid results (Valentine, 2009). The research reports identified for a particular review are likely to vary in terms of their adherence to a range of methodological indicators (Maggin, 2015). As such, it is often necessary to consider the methodological strength of the research base to properly contextualize the results. For instance, Maggin, Chafouleas, Johnson, and Ruberto (2012) found that the research examining the effects of group contingency

interventions on classroom behaviors often did not include measures of procedural fidelity. Providing evidence that the intervention was implemented as intended is critical for making claims that the intervention was in fact responsible for positive outcomes (Collier-Meek, Fallon, Sanetti, & Maggin, 2013). By collecting information on this particular methodological issue and others, the authors were able to provide a more nuanced overview of the evidence base aside from simply reporting the results. After all, the results of a study are meaningful but must be interpreted in relation to the research design and the various points where bias can be introduced. It is important to note that methodological strength is a multidimensional construct with many different features and that there are no perfect studies (Valentine, 2009). Moreover, the relevant methodological features for a given review will depend on the type of research being considered. Fortunately, researchers have developed methodological standards and related tools that can address a large range of different designs including group-based experimental designs, single-case research, measurement studies, and qualitative methods (e.g., Cook et al., 2014). We recommend that these quality indicators be used to evaluate and contextualize the strength of the findings within systematic reviews.

Describing participants and settings. The evaluation of study quality is important for establishing confidence in the research results; however, even the results of the best designed studies require consideration of the individuals and contexts to which the results can generalize (Valentine, 2009). Systematic reviewers are well positioned to provide a detailed overview of the characteristics of participants and settings addressed in a particular body of work. For example, a research team conducting a review on the effects of a reading comprehension intervention for adolescents might consider a range of student characteristics and/or settings in which the research has been conducted. Researchers can then ascertain the extent to which interventions work for particular students in particular settings, and whether gaps exist that can be addressed through future research.

Describing the variables under study. Systematic reviews are compelling because they organize information across several sources to draw conclusions that are more generalizable than those from a single study. This includes the variability inherent across research participants and contexts included in different studies as well as the variables and phenomena being investigated. As a result, those engaged in conducting systematic reviews should aim to describe the patterns in the outcomes and other key variables used across studies, including important moderator variables. Examples of important moderator variables for systematic reviews of interventions include fidelity of intervention implementation; duration, frequency, and intensity of intervention; and intervention cost (Cordray & Morphy, 2009). Examining variables across studies allows reviewers to ascertain what constructs researchers are studying. Looking across studies, reviewers can determine the uniformity of construct definitions. For example, reviewers might conduct a systematic investigation of variability in operational definitions and data collection procedures for the independent variables, dependent variables, and educational phenomena under study (Knowles, Meng, & Machalicek, 2015). Such variability will ultimately depend on key factors such as research questions, topic of study, and study measures. It is the reviewer's responsibility to develop an approach for capturing this variability in a systematic way, so as to describe it to readers of the review.

Data Analysis Plan

The goal of a plan for data analysis is to synthesize the results from the review; synthesis is the process of integrating findings from the studies to answer the research question (Gough et al., 2013). The data analysis plan goes hand in hand with the research question. Following the extraction of data, the investigator is ready to synthesize the results obtained across studies. The specifics of the data analysis plan will depend on the type of data extracted and the overall goals of the systematic review. For example, quantitative data can either be statistically combined using meta-analytic procedures or summarized in table form, reported separately for each study. The advantage of meta-analysis relates to the improved precision of the statistical estimates obtained from aggregating quantitative results across samples drawn from several studies. These procedures also allow investigators the opportunity to estimate the consistency of the estimates through heterogeneity analyses and determine whether and to what extent conceptually relevant variables explain the variability in statistical estimates (Cooper et al., 2009).

The systematic review quality indicators described in the preceding paragraphs represent widely regarded reporting standards (e.g., APA, 2010; Higgins & Green, 2011; Moher et al., 2015). To pilot the proposed standards and examine their reliability, as well as to evaluate the extent to which systematic reviews published in *Behavioral Disorders* adhered to these standards, we applied the standards to systematic reviews published in this journal from 2005 to 2016. As such, we sought to address the following research questions:

Research Question 1: To what extent have systematic reviews published in *Behavioral Disorders* been consistent with methodological quality indicators for systematic reviews?
Research Question 2: Are there areas of particular strength and areas in need of more attention?

In the following sections, we describe the procedures used to address these questions.

Method

Study Identification Procedures

The purpose of the current review was to identify systematic reviews and meta-analyses of research published in *Behavioral Disorders* from 2005 to 2016 and evaluate them on a range of established quality indicators for systematic review. Three individuals identified systematic reviews and meta-analyses (two held doctorate degrees in special education and the third was enrolled in a PhD program). These individuals are all study authors and each has experience conducting systematic reviews and meta-analyses. The following procedures were used to identify the pool of studies reviewed for the current article. First, the authors conducted a hand search of all copies of *Behavioral Disorders* published between January 2005 and May 2016. This hand search included the review of each title and abstract of papers published across these years to determine whether the article should be read in greater detail. The hand search resulted in the identification of 24 systematic reviews published over the aforementioned time period. Second, an electronic search of these journal issues was also conducted as an extra measure to ensure that all relevant papers were identified. This electronic search was conducted in ERIC and executed with the following Boolean phrase: pub("Behavioral Disorders") AND ("meta-analysis" OR "systematic review" OR "literature review") with a delimiter also applied to the time frame to ensure returns between the dates of interest. The electronic search returned a total of 309 papers with a majority of these published in other journals with titles containing the term "Behavioral Disorders." The research team therefore discarded those papers published in other journals and reviewed the titles and abstracts of the remaining papers. This process resulted in the identification of zero additional papers being included in the present review. Each of the 24 papers identified was then retrieved as a portable document format (PDF) and placed into an electronic file-sharing folder that all members of the research team could access.

Inclusion Criteria

The inclusion criteria used to select research reports for the present review were based on (a) the journal of publication and (b) the methodological procedures used within the studies. Broadly, studies were included if systematic review procedures were employed to address the research question. More specifically, systematic reviews were differentiated from other research approaches and less structured reviews if (a) the stated purpose of the research was to review extant literature and (b) there was a formal method section included in the article that described the process for selecting, searching, locating, and analyzing previously existing research studies. Regarding this latter criteria, literature reviews that were developed using narrative rather than systematic procedures were excluded from the present review. Moreover, there were no qualifications placed on the type of research that was included in the original review and, therefore, could include randomized or quasi-experimental, single-case, qualitative, and other research methods. Moreover, the search was restricted to systematic reviews and meta-analyses published between 2005 and 2016 to align with the publication of the Council for Exceptional Children (CEC) methodological quality indicators, which are widely viewed as a watershed development in the field, raising the standards for research (Odom et al., 2005).

Study Coding Procedures

We coded the identified set of systematic reviews and meta-analyses on 41 items or quality indicators in seven methodological domains drawn from established guidelines for the conduct of systematic reviews and meta-analyses from the APA (2010), *Cochrane Collaboration* (Higgins & Green, 2011), and the *What Works Clearinghouse* (WWC; 2013). Table 1 provides an overview of each of these methodological domains and the items used to operationalize them. The seven methodological domains include (a) the search procedures reported, which relate to the extent to which the authors were able to identify the universe of studies related to the research question; (b) the inclusion and exclusion criteria used within the review, which provide a basis for identifying studies that are relevant to the research questions; (c) the procedures used to retrieve studies deemed eligible, including explicit reference to relevant studies that could not be obtained; (d) the methods used to screen studies, which allows readers to evaluate whether the studies were consistently and accurately assessed for eligibility; (e) the procedures associated with implementing a coding scheme and related data collection processes; (f) the content of the coding scheme, which is essential for readers to determine the variables and issues that the research team considers most important in relation to the research question; and (g) the methods used to analyze the data, which encompasses both the approaches used to aggregate the descriptive results of the coding scheme and to synthesize the results across studies. All of the items were evaluated as either being present or not present unless otherwise noted.

Coder Training and Agreement

Two research assistants who had varying levels of expertise and experience with research methods conducted coding. The lead coder was a doctoral student in special education who had taken courses in research design and statistics and had experience in the conduct of previous systematic reviews. The secondary coder was also a special education doctoral student with no previous experience with systematic review procedures. Prior to initiating study coding, the research assistants underwent training that consisted of the following procedures: (a) an introduction to the study and the coding

Table 2. Interobserver agreement for each methodological domain sampled in the coding scheme.

Methodological domain	Total agreed	Total possible	Interobserver agreement (%)
Research question	7	7	100
Eligibility criteria	27	29	93
Search procedures	91	101	90
Retrieval procedures	24	29	83
Systematic screening	33	36	92
Coding scheme procedures	43	47	91
Coding scheme content	21	22	95
Data analysis plan	21	22	95
Total	267	293	91

Note. Interobserver agreement indices were computed based on independent coding of 30% of the studies which were randomly sampled; coder discrepancies were resolved through developing a consensus between coders.

protocol presented by the lead authors; (b) practice coding a subset of reviews and meta-analyses that would have been eligible for the study, except for their publication year (i.e., publication year prior to 2005); and (c) computing agreement statistics for the reviews used for coding practice until there was at least 90% agreement across coders for all codes. Coders were then given studies to code independently, with the lead coder assigned to review all of the eligible studies and the secondary coder assigned a random subset of 30% of the studies to establish interrater agreement. Interrater agreement was computed using a percentage agreement formula in which total agreements for each code were divided by the total agreements plus disagreements. Mean percentages across coders for each methodological domain are presented in Table 2.

Data Analysis Plan

To examine the extent to which systematic reviews and meta-analyses published in *Behavioral Disorders* from 2005 to 2016 are consistent with established quality indicators, we computed the proportion of systematic reviews that met each quality indicator. To illustrate the extent to which each domain and item was marked as present or not present, data were graphed using a horizontal stacked bar chart indicating the proportion of reviews that addressed the criterion or not. The data for each methodological domain were then sorted to illustrate those criteria that were addressed more often.

Results

Quality Indicator Application

Search procedures. The search procedure items provided an overview of the extent to which the included reviews were consistent with recommended strategies and practices for identifying relevant studies on the review topic. Results for each item are graphically depicted in Figure 1. With regard to this particular methodological domain, we identified a number of areas in which authors consistently described their procedures. For instance, all or nearly all of the reviews identified the databases searched to locate potentially eligible studies (n = 24) and provided a list of keywords that were used within these database searches (n = 22). Four additional criteria were determined to be present across the majority of included reviews, including the use of hand searches of journals to locate studies (n = 18), the examination of unpublished literature sources (n = 18), the review of citation lists of included studies (n = 16), and the process of reviewing titles and abstracts as part of the electronic search described (n = 14). The remaining eight criteria related to search procedures were found to be reported in fewer than 50% of the reviews and included examining the citation lists of previous reviews (n = 11); describing the qualifications of those charged with coding studies (n = 8); specifying language delimiters placed on searches (n = 7); reporting interobserver agreement on the search process (n = 5); consulting subject matter experts to verify the citation list of the review (n = 5); contacting authors of primary studies to determine whether there were any additional, unpublished reports (n = 4); specifying the time period or date of the search process (n = 4); and searching registries of completed or ongoing trials or studies to ensure all relevant research on the topic was obtained (n = 1).

Inclusion and exclusion criteria. The inclusion and exclusion criteria domain provided an overview of the extent to which the included reviews provided information necessary to clearly determine the population of studies that would be eligible for the review. Results for each item included in the domain are graphically presented in the upper left-hand panel of Figure 1. Two items were scored on a 3-point scale and were coded as either not present, partially present, or present. The first of these items pertained to the research questions and overall review purpose. Results indicated that a statement of the review's research questions or study purpose were present for most of the reviews (n = 19), whereas fewer were rated as partially present (n = 3) or not present (n = 1). With regard to whether the operational characteristics of the variables central to the review were adequately described, the majority of the reviews were rated as having met this criteria (n = 15) with the

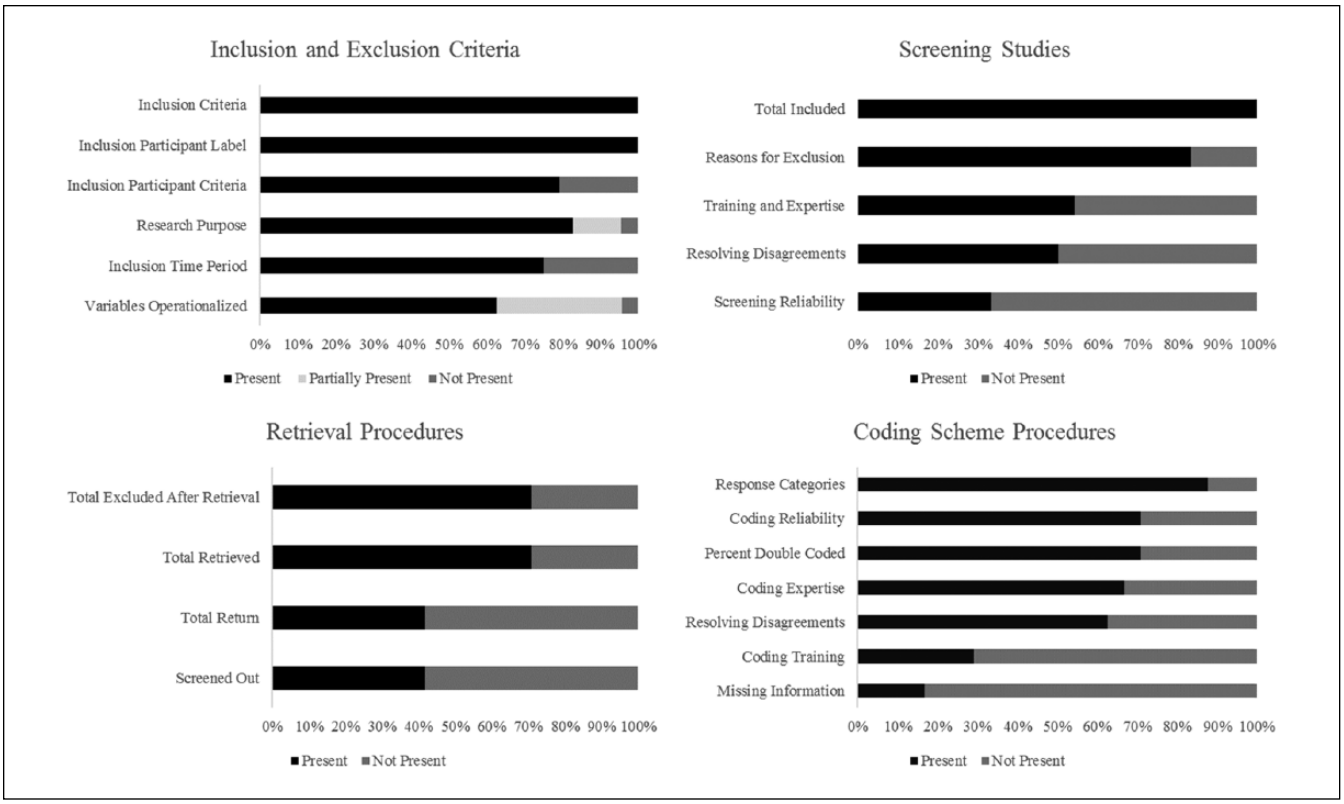

Figure 1. Percentage of systematic review quality indicators marked as present and not for the inclusion and exclusion criteria, screening studies, retrieval procedures, and coding scheme domains.

remaining reviews rated as either partially present (n = 8) or not present (n = 1). Regarding whether the particular diagnostic label for participants had been clearly reported, all of the reviews met this criterion (n = 24). Moreover, all of the reviews made clear the types of research designs that were eligible for review (n = 24). Approximately two thirds of the reviews were coded as describing the necessary characteristics of the student population eligible for the review (n = 19) and the time period for eligible studies (n = 18).

Retrieval procedures. These criteria evaluated reporting of the process used to narrow the initial pool of potentially eligible studies to those included in the review. Results for each item included in the domain are graphically presented in the lower left-hand panel of Figure 1. Application of the coding scheme indicated that fewer than half of the reviews provided information on the number of citations returned from the initial database search (n = 10) and the number of these citations that were excluded during the initial search (n = 10). Most of the reviews, however, contained information on the number of studies that were successfully retrieved (n = 17) and the number of studies retrieved and subsequently screened from further review (n = 17).

Screening studies. The process for screening studies domain provided an assessment of the extent to which the included reviews adhered to recommended practices for reporting on the screening process. Results for each item included in the domain are graphically presented in the upper right-hand panel of Figure 1. These results indicated that each of the reviews contained information pertaining to the total number of studies included in the review (n = 24) and that a majority of the reviews provided the reasons that those individual studies excluded from the review were screened from further review (n = 20). Slightly more than half of the reviews contained information on the credentials and training for those individuals charged with implementing the screening process (n = 13). Precisely half of the reviews included information about the process for resolving disagreements across those members of the research team engaged in the screening process (n = 12) with a third reporting reliability information on the screening procedures (n = 8).

Coding scheme procedures. The coding scheme procedures domain included items that assessed the methods used to systematically code the information within primary studies. Results for each item included in the domain are graphically presented in the lower right-hand panel of Figure 1. Most of the reviews described the response categories that coders could select from to code the relevant information from the studies (n = 21). Moreover, between two thirds and three quarters of the reviews described the expertise of

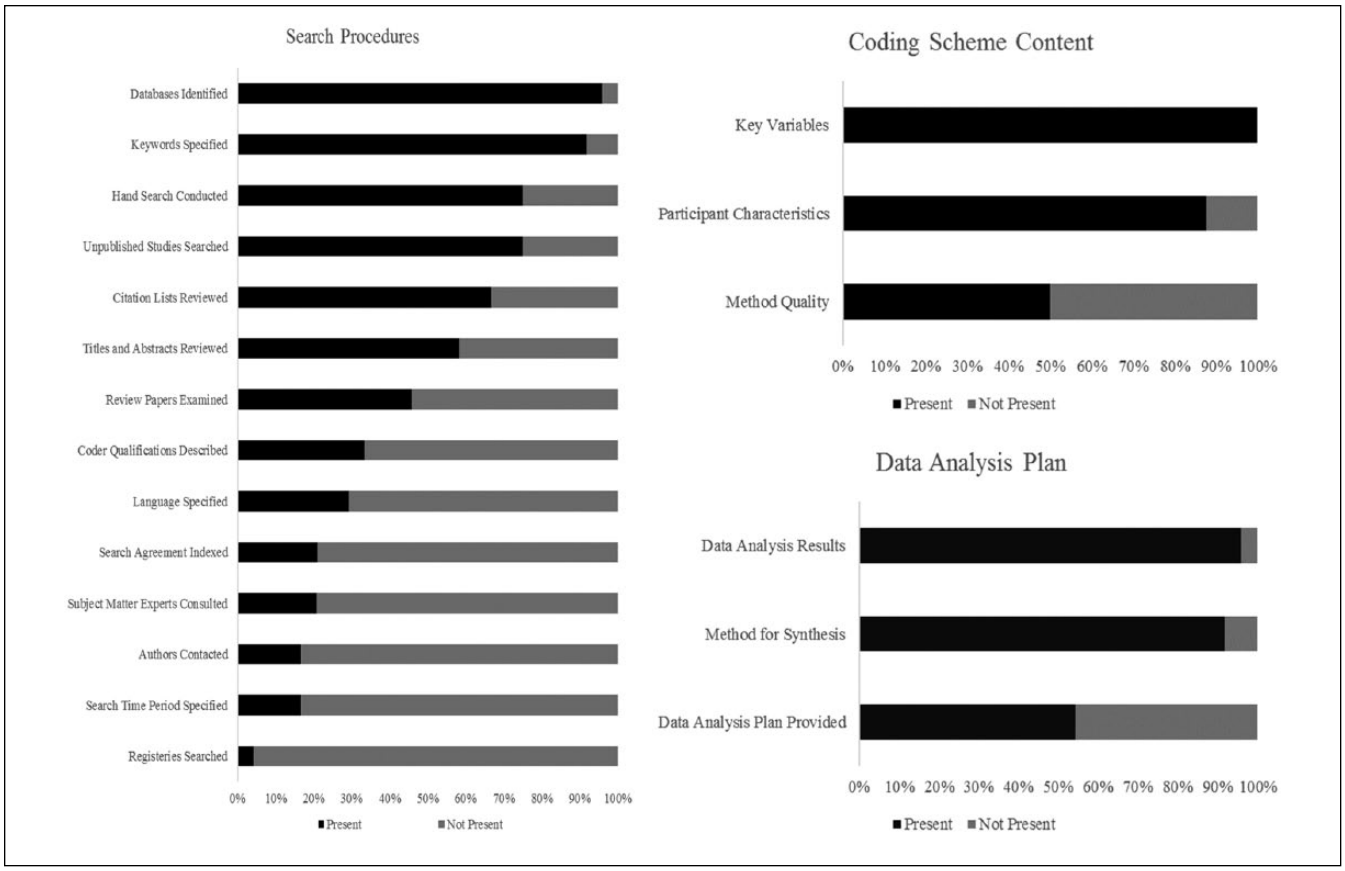

Figure 2. Percentage of systematic review quality indicators marked as present and not for the search procedures, coding categories, and data analysis plan domains.

those coding the studies (n = 16), the proportion of studies that were double coded for reliability (n = 17), and the interobserver agreement of the coding scheme's application to the studies included in the review (n = 17). Fewer than two thirds of the studies provided information about the processes for resolving disagreements across coders (n = 15), the methods used to train and prepare coders to implement the scheme (n = 7), and the process used to address circumstances where key information was missing from the report altogether (n = 4).

Coding scheme content. The content of the coding schemes used across the included reviews was also considered. Results for each item included in the domain are graphically presented in the upper right-hand panel of Figure 2. Because the content of each coding scheme should be developed to address the particular research questions and study purposes of the review, these items focused on the general categories contained within the reviews. It was determined that each of the coding schemes explicitly collected information on the primary variables of interest and the majority included information pertaining to the participants of the study (n = 21). Exactly half of the studies considered issues of methodological quality (n = 12).

Data analysis plan. The data analysis plan items related to whether the reviews contained information on the processes and procedures that were used to analyze the data, such as the particular statistical models in a meta-analysis, the process for identifying themes and patterns within metasyntheses, or the use of descriptive statistics in cases where synthesis was not the goal of the review. Results for each item included in the domain are graphically presented in the lower right-hand panel of Figure 2. As with the coding scheme content, the particular data analysis plan was not the focus but rather the more general aspects of a data analysis plan. We found that most of the reviews did include information on the processes used to aggregate results across studies (n = 22) and identify patterns within the data on which to draw conclusions (n = 23). However, only slightly more than half provided a formal data analysis section (n = 13).

Discussion

The purpose of this article was to provide an overview and rationale for rigorous systematic review methods to ensure that the information drawn from these reviews is the most valid and current available. To examine the extent to which systematic reviews published in *Behavioral Disorders* were

consistent with the proposed standards, we applied the quality indicators to reviews published in the journal from 2005 to 2016. Findings indicated several areas of distinct strength with additional areas in need of more attention among the 24 reviews. Before discussing the specific findings of our review, it is worth briefly considering the importance that systematic reviews have for stakeholders in the fields of special education and EBD. As the field continues to place more emphasis on the importance of research in guiding the selection of practices and policies, it is critical that stakeholders have access to the most recent and accurate information on which to make decisions. The systematic review ultimately represents the most valid and reliable vehicle for examining and synthesizing research findings across studies to inform the development of empirically grounded sets of practices and policies. As such, it is necessary to take stock of the methods being used to identify both strengths and weaknesses to ensure that the field is producing the most dependable information to improve the outcomes and lives of students with EBD. In the following sections, we describe specific areas of strength and areas in which the field might improve.

Quality Indicator Application

The application of the review indicators led to the identification of several areas of strength across the included reviews. For instance, there were a number of specific methodological criteria that were present across all of the reviews. Examples include the specification of inclusion criteria, noting the particular special educational label necessary for inclusion, and providing information on the total number of studies in the review. The methodological domain with the greatest adherence across the reviews was the application of inclusion and exclusion criteria. As such, it appears that systematic review authors have effectively established the boundaries of the review for their intended audiences. In fact, many of the criteria that were consistently implemented across the reviews reflected procedural components that could be replicated by other researchers. As others have pointed out (Travers, Cook, Therrien, & Coyne, 2016), the notion of replicability is central to the scientific endeavor and is mandatory for developing research-based practices, programs, and policies. Within the context of systematic reviews, providing enough procedural information to replicate a systematic review allows others to evaluate the strength of the research and to confirm or challenge the validity of the research process.

The consistent adherence on the part of reviewers in providing detailed descriptions of the review procedures is certainly encouraging. Of course, there were some procedural areas where more consistent reporting is needed, such as describing the process for training coders on the coding scheme, providing details on the process for resolving coding disagreements, and addressing the methodological quality of the included studies. An examination of the quality indicators that were not consistently implemented reveals that many are related to the need to better report the results of the procedural indicators. For instance, the majority of the studies did not provide information on the reliability of the screening process, the total number of citations returned in the electronic search, and the number of total studies screened out during the search process. Rather than relate to issues of the reproducibility of the procedures, these indicators relate more to issues of transparency of the review. Research transparency is a multidimensional construct and is central to developing and disseminating empirically based information (Cook, 2014). In the current context, transparency refers to the process of providing specific information that would assist others in understanding and replicating the review process. Reporting the number of studies identified through a particular database, for example, allows others to verify the results of the search process. The information that appears to be missing most consistently across the sampled reviews, therefore, would allow others to empirically corroborate the review procedures. Including this information seems to be a worthy advance for review authors to strive toward, as our field continues to pursue the use of the most rigorous research methods at our disposal.

Concluding Remarks

The dissemination of rigorous research designed to address the complex and varied needs of students with EBD is an essential mission of *Behavioral Disorders*. In the present article, we outlined a series of quality indicators for conducting rigorous systematic reviews and applied these indicators to reviews published in *Behavioral Disorders* between 2005 and 2016. The results of this application led to the identification of several areas of methodological strength and some areas that have tended to be overlooked. Researchers engaged in conducting systematic reviews to inform practice and policy related to students with EBD are encouraged to use the current pilot standards as a guide to ensure the most rigorous systematic review methods are used to disseminate the most valid and reliable information. We conclude with some considerations for the foregoing research.

It is instructive to consider some of the challenges associated with conducting a rigorous systematic review. Among the primary challenges in our discipline is the tendency of systematic reviewers to generalize the results too broadly and to assume that the results apply to all students with EBD. It is our observation that researchers conducting a systematic review will often not describe the conceptual reasons a particular intervention or program is expected to work and for whom (Maggin, Zurheide, Pickett, & Baillie, 2015). Fortunately, there are published examples in which these conceptual links are brought to the forefront. For instance, Bruhn, McDaniel, and Kreigh (2015) conducted a review examining the empirical research for self-monitoring interventions for students with behavior problems. Rather than group all

self-monitoring interventions together and assume the same components, these authors examined the underlying practices used in each study and their connection to issues of behavioral function. As a result, the authors were able to provide clear and direct guidance to the reader on whether self-monitoring interventions worked, for whom, and under what conditions. It is worth noting that the process of integrating theoretical, contextual, and methodological nuances across studies is not an easy endeavor. However, it is incumbent on the review authors to make these connections explicit for the reader. More attention to the sample characteristics and conceptual underpinnings of the practices and policies identified as evidence-based by systematic reviewers will allow for more targeted use of effective methods.

It is also important to discuss some of the challenges of applying quality indicators to bodies of research. To be sure, assessing methodological quality is an important endeavor and one that is necessary to accurately contextualize research reports (Valentine, 2009). Despite these advantages, there is also a need to constructively use sets of quality indicators with the understanding that these are essentially research instruments that are subject to the same issues of validity and reliability as those quality indicators used within primary research. As such, additional vetting of the quality indicators forwarded here and their application to special education research is needed. Fortunately, recent research has provided a template of how such a process might be pursued (Cook et al., 2014). Finally, it may be tempting to assume that systematic reviews that address fewer indicators are of lower quality. In many instances, this might be the case although there might also be examples in which the particular research being reviewed or the focus of the review might preclude the implementation of all criteria. It is our recommendation, therefore, that future applications of these and other indicators be used more to describe the current status of research in a particular area rather than be used for evaluative purposes.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

American Psychological Association. (2010). *Meta-analysis reporting standards*. Washington, DC: Author.

Baumeister, R. F., & Leary, M. R. (1997). Writing a narrative literature review. *Review of General Psychology*, *1*, 311–320.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley.

Bruhn, A., McDaniel, S., & Kreigh, C. (2015). Self-monitoring interventions for students with behavior problems: A systematic review of current research. *Behavioral Disorders*, *40*, 102–121.

Collier-Meek, M. A., Fallon, L. M., Sanetti, L. M. H., & Maggin, D. M. (2013). Focus on implementation: Assessing and promoting treatment fidelity. *Teaching Exceptional Children*, *45*, 52–59.

Cook, B. G. (2014). A call for examining replication and bias in special education research. *Remedial and Special Education*, *35*, 233–246. doi:10.1177/0741932514528995

Cook, B. G., Buysse, V., Klingner, J., Landrum, T., McWilliam, R., Tankersley, M., & Test, D. (2014). Council for Exceptional Children: Standards for evidence-based practices in special education. *Teaching Exceptional Children*, *46*, 206–212. doi:10.1177/0040059914531389

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*, 365–383. doi:10.1177/001440290907500306

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–16). New York, NY: Russell Sage Foundation.

Cordray, D. S., & Morphy, P. (2009). Research synthesis and public policy. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 473–493). New York, NY: Russell Sage Foundation.

Gage, N., Gersten, R., Sugai, G., & Newman-Gonchar, R. (2013). Disproportionality of English learners with emotional and/or behavioral disorders: A comparative meta-analysis with English learners with learning disabilities. *Behavioral Disorders*, *38*, 123–136.

Gersten, R., Fuchs, L. S., Compton, D., 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.

Gough, D., Oliver, S., & Thomas, J. (2013). *Learning from research: Systematic reviews for informing policy decisions: A quick guide*. London, England: Alliance for Useful Evidence.

Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 5.1). Chichester, UK: Wiley-Blackwell.

Knowles, C., Meng, P., & Machalicek, W. (2015). Task sequencing for students with emotional and behavioral disorders: A systematic review. *Behavior Modification*, *39*, 136–166.

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

Maggin, D. M. (2015). Considering generality in the systematic review and meta-analysis of single-case research: A response to Hitchcock et al. *Journal of Behavioral Education*, *24*, 470–482.

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*, 3–8.

Maggin, D. M., Wehby, J. H., Farmer, T. W., & Brooks, D. S. (2016). Intensive interventions for students with emotional and behavioral disorders: Issues, theory, and future Directions. *Journal of Emotional and Behavioral Disorders*, *24*, 127–137.

Maggin, D. M., Zurheide, J., Pickett, K. C., & Baillie, S. J. (2015). A systematic evidence review of the check-in/check-out program for reducing student challenging behaviors. *Journal of Positive Behavior Interventions*, *17*, 197–208.

Maggin, D. M., Johnson, A. H., Chafouleas, S. M., Ruberto, L. M., & Berggren, M. (2012). A systematic evidence review

of school-based group contingency interventions for students with challenging behavior. *Journal of School Psychology*, *50*, 625–654.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*, Article 1.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*, 137–148.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academies Press.

Talbott, E., Maggin, D. M., Van Acker, E. Y., & Kumm, S. (in press). Quality indicators for reviews of research in special education. *Exceptionality*.

Travers, J. C., Cook, B. G., Therrien, W. J., & Coyne, M. C. (2016). Replication in research in special education. *Remedial and Special Education*, *27*, 195–204.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 129–146). New York: Russell Sage Foundation.

What Works Clearinghouse. (2013). *Procedures and standards handbook* (Version 3.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf