5-2023

Evidence-Based Assessment in Special Education Research: Advancing the Use of Evidence in Assessment Tools and Empirical Processes

Elizabeth Talbott
William & Mary, ehtalbott@wm.edu

Andres De Los Reyes

Devin M. Kearns

Jeannette Mancilla-Martinez

Mo Wang

Recommended Citation

Talbott, Elizabeth; De Los Reyes, Andres; Kearns, Devin M.; Mancilla-Martinez, Jeannette; and Wang, Mo, Evidence-Based Assessment in Special Education Research: Advancing the Use of Evidence in Assessment Tools and Empirical Processes (2023). *Exceptional Children*, 89(4), 467-487. https://doi.org/10.1177/0014402923117109

Policy Paper

Evidence-Based Assessment in Special Education Research: Advancing the Use of Evidence in Assessment Tools and Empirical Processes

Exceptional Children
2023, Vol. 89(4) 467-487

DOI: 10.1177/00144029231171092
journals.sagepub.com/home/ecx
Sage

Elizabeth Talbott[1], Andres De Los Reyes[2], Devin M. Kearns[3], Jeannette Mancilla-Martinez[4], and Mo Wang[5]

Abstract

Evidence-based assessment (EBA) requires that investigators employ scientific theories and research findings to guide decisions about what domains to measure, how and when to measure them, and how to make decisions and interpret results. To implement EBA, investigators need high-quality assessment tools along with evidence-based processes. We advance EBA in three sections in this article. First, we describe an empirically grounded framework, the Operations Triad Model (OTM), to inform EBA decision-making in the articulation of relevant educational theory. Originally designed for interpreting mental health assessments, we describe features of the OTM that facilitate its fusion with educational theory, namely its falsifiability. In turn, we cite evidence to support the OTM's ability to inform hypothesis generation and testing, study design, instrument selection, and measurement validation. Second, we describe quality indicators for interpreting psychometric data about measurement tools, which informs both the development and selection of measures and the process of measurement validation. Third, we apply the OTM and EBA to research in special education in two contexts: (a) empirical research for causal explanation and (b) implementation science research. We provide open data resources to advance measurement validation and conclude with future directions for research.

Sound assessment evidence comprises the bedrock on which we build foundational principles of evidence-based practices (EBPs) in special education (De Los Reyes, Talbott, et al., 2022). Evidence-based assessment (EBA) requires that investigators employ scientific theories and empirical findings to guide decisions about what domains to measure, how and when to measure them, and how to leverage measurement outcomes to make informed decisions about service delivery (e.g., intervention planning, estimating intervention responses; see Hunsley & Mash, 2007; Youngstrom et al., 2017). EBA plays a key role in special education research, as illustrated by the evidence-based processes that characterize curriculum-based research. These include curriculum-based *measurement*, which requires the use of precise,

[1]William and Mary
[2]University of Maryland, College Park
[3]University of Connecticut
[4]Vanderbilt University
[5]University of Florida

Corresponding Author:
Elizabeth Talbott, Department of Curriculum and Instruction, School of Education, William and Mary, Williamsburg, VA 23185, USA.
Email: ehtalbott@wm.edu

standardized tools to measure accuracy and speed in academic skills (Deno, 1985; Hosp et al., 2016); curriculum-based *assessment*, which requires a review of results using those tools alongside a review of extant data and contexts for measurement (Hosp et al., 2014); and curriculum-based *evaluation*, which is a broad, systematic approach to using multiple sources of data collected from tools meeting psychometric thresholds in the context of knowledge about how children learn (American Educational Research Association [AERA], 2014; Hosp et al., 2014). Similarly, scholars in implementation science leverage EBA principles to develop tools that estimate intervention fidelity in context, with a goal of promoting the use of evidence in routine practice across diverse settings (Cook & Odom, 2013; Odom et al., 2020). EBA thus entails a comprehensive approach to the application of both theory and psychometrics in the education sciences in order to inform EBPs, resulting in an integrated approach to the use of evidence (Podsakoff et al., 2012; Youngstrom et al., 2017). EBA includes evidence-based tools for measurement, processes germane to testing the interpretation of scores and their relevance according to theory and research (i.e., measurement validation), and the implementation of those tools with diverse stakeholders in applied settings (AERA, 2014; Borsboom et al., 2004; Hunsley & Mash, 2007; Shear & Zumbo, 2014). Decision-making in assessment is central to the EBA process, equal in importance to the development and application of tools with strong psychometric characteristics (AERA, 2014).

Scholars in special education research share a core set of "best practices" in assessment, namely, use of multimethod, multi-source, multidomain approaches to understand, measure, collect, and interpret data (e.g., Gersten et al., 2005; Hosp et al., 2014; Hunsley & Mash, 2007; Individuals With Disabilities Education Improvement Act, 2004). This approach addresses challenges with educating students with disabilities, given their diverse needs (e.g., social, emotional, and behavioral health; academic performance; *Endrew F. v. Douglas County School District*, 2017). Within this approach, investigators gather, interpret, and use evidence from multiple data sources, including parents, teachers, trained observers, and youth (De Los Reyes, 2011) as well as multiple, diverse measures of academic achievement (AERA, 2014).

Within EBP in special education and allied fields (e.g., medicine, education, clinical and school psychology), research in assessment has not kept pace with research on interventions (Jensen-Doss, 2011; Jensen-Doss & Hawley, 2010; Straus et al., 2019). Driven by federal law and policy, researchers in special education have devoted considerable attention to research in EBPs, establishing quality indicators and addressing their application to implementation science (Cook & Odom, 2013; Odom et al., 2020). The Council for Exceptional Children (CEC) has established standards for EBPs in special education (CEC, 2014), with researchers having previously established guidelines for determining them (Cook et al., 2009). Several of these pieces (CEC, 2014; Cook et al., 2009; Gersten et al., 2005; Thompson et al., 2005) have referenced psychometric criteria for evidence-based assessment tools. However, none have outlined a comprehensive approach to employing EBA processes for special education. This approach ought to include not only conceptually grounded and evidence-based guidelines relevant to developing and selecting psychometrically robust tools (Podsakoff et al., 2012), but also practices relevant to interpreting and integrating the patterns of scores produced by these tools (De Los Reyes & Epkins, 2023).

Research in assessment has not kept pace with research on interventions.

In the absence of empirical guidelines from the field, government agencies such as the U.S. Institute of Education Sciences (IES) have stepped in to fill the void, particularly with regard to providing guidance in the identification of outcome measures (see Schneider, 2020; U.S. Department of Education, 2022a, 2022b). In its Standards for Excellence in

Education Research (SEER), the IES has issued broad guidance that, on the surface, appears to be well aligned with EBA, advising researchers to "examine both the immediate impact of their intervention on outcomes of interest as well as its impact on relevant distal outcomes and the potential that initial impacts may fade over time" (U.S. Department of Education, 2022a). Yet, the director of the IES (Schneider, 2020), the SEER standards (U.S. Department of Education, 2022a), and the What Works Clearinghouse (WWC; U.S. Department of Education, 2022b) have issued narrow guidelines, advising education researchers to use "*common outcome measures* with name recognition" to determine the effectiveness of interventions (Schneider, 2020) and avoid selecting outcome measures that are "overaligned" with interventions (U.S. Department of Education, 2022a, 2022b). This narrow piece of guidance from the IES appears to have been issued outside of the broader context of EBA, wherein appropriate outcome measures would be identified in the context of theory on the EBA decision-making process as well as findings from EBA-informed measurement validation studies (AERA, 2014; Borsboom et al., 2004; Hunsley & Mash, 2007; Kazdin, 2016; Podsakoff et al., 2012; Shear & Zumbo, 2014). That is, the validity of a given instrument does not exist in isolation but instead resides squarely in the measurement context, which includes the purpose of assessment; such as screening, diagnosis, progress monitoring, and intervention decision-making (AERA, 2014; Youngstrom et al., 2017). Thus, evaluating tools for their proposed uses with empirical validation must play a central role in the EBA process in education research (AERA, 2014; Kazdin, 2016). Importantly, beyond this guidance, the field lacks procedures for interpreting and integrating data produced across tools used in comprehensive assessments (see also Beidas et al., 2015; De Los Reyes et al., 2019).

To address this crucial need for guidance, we provide conceptual and empirical direction for EBA in special education research. Our guidance is framed by research in the education (AERA, 2014; U.S. Department of Education, 2013, 2022a, 2022b) and psychological sciences (American Psychological Association, 2021; De Los Reyes & Langer, 2018; Kazdin, 2016; Hunsley & Mash, 2007; Youngstrom et al., 2017) and is presented in three sections. First, we describe an empirically grounded framework, the Operations Triad Model (OTM), to inform EBA decision-making in the articulation of relevant educational theory. We describe features of the OTM, which was originally designed for interpreting mental health assessments, that facilitate its fusion with educational theory, namely its falsifiability. In turn, we cite evidence to support the OTM's ability to inform hypothesis generation and testing, study design, instrument selection, and measurement validation. Second, we describe quality indicators for interpreting psychometric data about measurement tools, which informs both the development and selection of measures and the process of measurement validation. Third, we apply the OTM and EBA to research in special education in two empirical contexts: (a) empirical research for causal explanation and (b) implementation science research. We provide open data resources to advance measurement validation research and conclude by providing future directions for research.

To address this crucial need for guidance, we provide conceptual and empirical direction for EBA in special education research.

Conceptual and Empirical Framework for EBA

As a framework for interpreting multivariate patterns of measurement outcomes, the OTM's value lies in two characteristics. The first is its falsifiability. The second is its flexibility for guiding the EBA process in the context of relevant educational theory and in a way that traverses diverse areas of special education research (e.g., language, reading, writing, mathematics, executive functioning, and mental health; see Table A1 in the online supplemental materials). By design, the multivariate nature of assessment in education presents various complexities with using data to make precise, accurate decisions relevant to student learning and learning outcomes. Yet, investigators have little guidance about how to interpret and integrate data derived from this approach (Achenbach, 2020; Beidas et al., 2015). Consequently, many investigators currently use analytic models that are inconsistent with an emerging body of work on use of this approach across the social sciences, including in education (De Los Reyes et al., 2019; De Los Reyes, Talbott, et al., 2022; De Los Reyes, Tyrell, et al., 2022; Talbott et al., 2021). That is, since the 1950s, researchers have sought convergence in the data they collect to measure diverse domains using multiple methods and data sources (i.e., *converging operations*; Campbell & Fiske, 1959; Garner et al., 1956). A key assumption underlying an emphasis on convergence is that when data do not converge—when data produce discrepant estimates—this signals instances in which data lack meaning and thus cannot inform our understanding of education-relevant domains (see also De Los Reyes, Tyrell, et al., 2022). This is a key assumption, particularly in light of the data conditions that typify education research.

For example, data from studies conducted over the past 60 years of educational and psychological research make clear that multiple data sources (e.g., survey reports from parent, teacher, and youth informants; systematic direct observations; academic achievement tests; researcher-designed outcome measures; curriculum-based measures) are likely to provide conflicting estimates of child and adolescent behavior and academic performance. As evidence of this notion, we present in Table A1 results of meta-analytic reviews across a host of education-relevant domains. The data sources used to estimate these domains include multi-informant ratings of youth behavior as well as youth scores on various outcome measures (i.e., language, reading, writing, mathematics, executive functioning, and mental health). Across these meta-analyses, correspondence among scores taken from the various data sources (e.g., as indexed by Pearson r correlations) range from nearly 0 to nearly 1.00 (see Table A1). Importantly, two characteristics of the research on these low correspondence levels among sources indicate that they contain data that actually inform our understanding of education-relevant domains. First, low correspondence among sources cannot be explained away by the lack of psychometric rigor inherent in the instruments used to collect data. In fact, low correspondence occurs even when measures used are psychometrically sound and meet quality indicators. This suggests that measurement confounds—irrelevant variance in score estimates, such as rater bias or random error—do not fully explain low correspondence (Achenbach, 2011; Dirks et al., 2012).

Second, over 15 years of carefully controlled studies demonstrate the discrepant estimates revealed in low correspondence rates often reflect domain-relevant information. By "domain-relevant information," we mean meaningful variations in youth behavior and performance and thus data that inform our understanding of measured domains (De Los Reyes, Talbott, et al., 2022; De Los Reyes, Tyrell, et al., 2022; Lerner et al., 2017; Talbott & De Los Reyes, 2022). For example, in the area of measurement in early reading, researchers frequently combine the two foundational skills of word recognition (word pronunciation and decoding) when in fact the two are independent domains, particularly among readers with dyslexia (Castles & Coltheart, 1993; Kearns et al., 2019).

Despite these findings, this emerging work on discrepant estimates has yet to meaningfully inform education research and theory. As evidence of this, consider that to this day, researchers often "reconcile" discrepant estimates using analytic models (e.g., combinational algorithms, composite scores, primary and secondary outcome measures; structural models) or measurement techniques (e.g., tests of measurement invariance) that emphasize converging data and assume the discrepancies reflect measurement confounds (e.g., see Bauer et al., 2013; De Los Reyes, 2011; De Los Reyes & Epkins, 2023; Martel et al., 2021; Olino et al., 2018; Watts et al., 2022).

This historic emphasis on convergence in assessment (e.g., the IES's recent call for the use of common outcome measures) has created barriers to the advancement of EBA. In particular, we see an absence of guidance to researchers who seek an "appropriate balance between measures closely aligned with an intervention and those of generalized performance" (Gersten et al., 2005, p. 151), particularly for those who seek to select measures designed to assess near transfer, mid-transfer, and far transfer of skills (Clemens & D. Fuchs, 2022). In fact, this emphasis on convergence appears misaligned with an individualized, precise approach to EBA and EBP, long-standing hallmarks of special education research (i.e., White, 1986). In these respects, the emphasis on convergence is likely to create barriers in implementation of EBA and EBP with diverse communities of learners if (a) the goal is for all students to achieve a single, elusive outcome on a particular assessment and (b) EBA and EBP are not adapted and delivered in partnership with key stakeholders (see also Baumann & Cabassa, 2020; von der Embse & De Los Reyes, 2023).

Taken together, the history of work on discrepant estimates involves researchers characterizing all of these discrepancies as measurement confounds and thus barriers to sound decision-making. Yet, the empirical work suggests otherwise—that there are times when these discrepant estimates contain valuable data. Education researchers require an approach to empirically detect these exact forms of discrepant estimates—those that reveal opportunities to improve the utility of assessment within the systematic intervention process (De Los Reyes, Talbott, et al., 2022; Talbott & De Los Reyes, 2022). To this end, we describe the OTM, a conceptual and empirical approach that allows researchers to distinguish domain-relevant discrepant estimates from those discrepancies that reflect measurement confounds.

OTM

As an approach to EBA, the OTM articulates three combinations of multivariate patterns of findings that commonly occur in social science research accompanied by their underlying meaning (De Los Reyes et al., 2013; see Figure A1 in the online supplement illustrating the OTM). The OTM is designed to aid investigators as they articulate the relevant educational theory and identify prior empirical research that drives the development of their research hypotheses and selection of measures. This relevant theory should also guide the measurement validation process (AERA, 2014; Kazdin, 2016; Podsakoff et al., 2012), which is closely associated with the quality indicators for measurement tools in Table A2 of the online supplemental materials. We have provided a graphical depiction of the linkages among the OTM, relevant theory, quality indicators, and measurement validation research in Figure A2. Evidence of the considerable heuristic value of the OTM comes from prior applications of the framework and the corresponding study of domain relevance in the discrepant estimates produced within a variety of assessment scenarios. This work represents a diverse set of areas, including (a) youth mental health (De Los Reyes et al., 2015), (b) special education (De Los Reyes et al., 2019; Talbott & De Los Reyes, 2022), (c) attention deficit hyperactivity disorder (Tamm et al., 2021), (d) autism spectrum disorder (ASD; Lerner et al., 2017), (e) implementation science (De Los Reyes, Talbott, et al., 2022), (f) physiological functioning (De Los Reyes et al., 2020; De Los Reyes & Aldao, 2015), (g) family functioning (De Los Reyes & Ohannessian, 2016), and (h) adult mental health (De Los Reyes & Makol, 2022). The OTM facilitates interpretation of results when data sources agree (converging operations) or disagree (diverging operations or compensating operations). The principle of converging operations—if not the term itself—describes many researchers' orientation toward interpretation of multimodal measures. Researchers with a converging-operations orientation test hypotheses (e.g., whether a special education intervention program is effective) and evaluate them using multiple measures aligned with the hypothesis, with the understanding that all measures used should point to the hypothesized conclusion (e.g., each source used to estimate outcomes indicates that the intervention is efficacious).

If researchers adhere to the principle of converging operations, they are likely to treat discrepant findings as evidence against the value of their hypotheses or about the quality of their measures (e.g., random error, rater bias; Watts et al., 2022). The concern about such interpretations is that discrepant patterns of performance on multiple methodologically distinct measures within the same domain may provide information about the relation between domain-relevant learner characteristics (e.g., the contingencies that elicit their needs for services) and performance on the assessment or response to instruction. In short, converging-operations interpretations provide an incomplete account of all possible multivariate outcomes, particularly in research areas where discrepant outcomes commonly occur (see Table A1).

Within the OTM, there is an understanding that not only might discrepant outcomes commonly occur but they also are not all created equal. Diverging operations characterizes those discrepant estimates that reflect domain-relevant phenomena, such as the measured domain varies across contexts or reflects the idea that social environments vary in their capacities to influence the behaviors being measured (e.g., parent report of home behavior vs. teacher report of school behavior). Conceptually, discrepant estimates reflecting diverging operations are distinguishable from compensating operations. Compensating operations reflect discrepant estimates that stem from artifactual characteristics of the measurement process, including (a) measurement confounds noted previously (e.g., random error, rater biases), (b) tools used to assess the domain do not meet adequate psychometric standards (AERA, 2014), (c) high-quality tools have been administered incorrectly, and (d) current findings indicate the need for additional research with the use of a particular high-quality tool.

Falsifiability of the OTM and Empirical Support. The OTM conceptualizes discrepant estimates using concepts beyond that of converging operations. In this respect, it is important to consider a key idea, namely, that the OTM's notions about discrepant estimates are, by design, falsifiable. That is, diverging operations and compensating operations are distinguishable in that each concept characterizes discrepant estimates in fundamentally different ways. If one concept characterizes discrepant estimates as relevant, whereas another concept characterizes these same discrepancies as irrelevant, then the means for empirically adjudicating this distinction falls squarely within the purview of measurement validation. A core principle of measurement validation involves testing relations between scores taken from instruments undergoing evaluation and scores taken from well-established, domain-relevant validity criteria. If an index of discrepant estimates relates to a domain-relevant validity criterion, then it can no longer be considered irrelevant. Particularly powerful demonstrations come from not only demonstrating these links between discrepant estimates and validity criteria but also doing so when using criteria that are untethered in instrumentation or in a way that avoids shared method biases (see De Los Reyes et al., 2023).

Along these lines, there exists a wealth of empirical support for the OTM's contention that not all discrepant estimates are created equal. We graphically depict this support in Figure A3 in the online supplemental materials. As depicted in Figure A3, several characteristics of this evidence enhance the interpretability of the findings. In particular, studies have demonstrated links between discrepant estimates and criterion variables using multiple measurement modalities. This rules out the possibility that shared-method biases explain the findings (see Garb, 2003). Further, links between discrepant estimates and domain-relevant validity criteria exist across multiple education-relevant domains, including psychosocial difficulties and strengths, as well as areas of psychosocial impairment or life interference (e.g., parenting, family relationships). In fact, this empirical support recently culminated in an extension of the OTM—classifying observations necessitates theory, epistemology, and testing (CONTEXT)—that instantiates its key principles in a measurement validation paradigm. As a validation paradigm, CONTEXT advances beyond traditional paradigms (i.e., Campbell & Fiske, 1959) in that it guides the construction of validation studies that are capable of detecting domain-relevant data when findings converge (i.e., converging operations) but also when they meaningfully diverge (i.e., diverging operations). Recent work provides a full description of this paradigm (De Los Reyes et al., 2023).

Domains to Probe for Empirical Evidence of Diverging and Compensating Operations

The falsifiability of the OTM lends itself well to use in hypothesis generation and testing across special education. Along these lines, we highlight two such areas for further exploration, one in which the discrepant estimates may very well reflect diverging operations and one in which the discrepant estimates might be parsimoniously explained by compensating operations.

> *The falsifiability of the OTM lends itself well to use in hypothesis generation and testing across special education.*

Empirical Example of Diverging Operations. Despite an emphasis on convergence, discrepant estimates are prevalent across domains in special education research (as seen in Table A1) and are thereby worthy examples for empirical exploration. For example, in a meta-analysis, Garcia and Cain (2014) identified nearly 700 correlations between measures of decoding and reading comprehension ranging from 0 to .96. In exploring the domain relevance of these discrepant estimates, they found that student age and listening comprehension skills were significant moderators of the relationship. In addition, characteristics of the measurement process also acted as significant moderators, including how decoding was assessed (i.e., whether students had help with decoding or the text was read aloud; Garcia & Cain, 2014). Their analysis of these discrepant estimates indicated both domain relevance (i.e., within the simple view of reading; Hoover & Gough, 1990) and a clear contribution of measurement characteristics in assessing those domains (Garcia & Cain, 2014).

Empirical Example of Compensating Operations. If assessment designs are not conducive to meaningfully interpreting discrepancies, then compensating operations might apply, namely, discrepant estimates that reflect measurement confounds. Ledford and colleagues (2015) provide an example of these confounds. In this study, the authors analyzed variation in findings in single case design research (SCD) associated with different interval-based recording systems employed to estimate behavior: (a) momentary time sampling, (b) partial interval recording (PIR), and (c) whole interval recording (WIR). Ledford and colleagues identified potential measurement confounds in the application of these methods, where PIR was likely to overestimate the occurrence of behavior, even as WIR methods were likely to underestimate actual behavior occurrence (Ledford et al., 2015). The result of these particular measurement confounds may be an overestimate of intervention effects, leading to Type I errors in the identification of EBPs (Ledford et al., 2015).

Empirical Solutions Within the OTM. In highlighting research domains in special education that may exemplify features of the OTM, a question arises: Which research practices or study designs may be conducive to putting the ideas discussed previously to the test? Here, we highlight a few promising paradigms to consider. When empirical findings diverge or are expected to diverge based on previous research, scientists can generate and test hypotheses about whether these results represent diverging operations or compensating operations. Furthermore, scientists can use the open science tool of preregistration to prepare and share their study hypotheses in advance, anticipating which findings are likely to converge or diverge prior to empirical testing (see Johnson & Cook, 2019; Lombardi et al., 2023). Thus, we describe empirical approaches that education scientists can use to test specific hypotheses within the OTM (i.e., through SCD) and determine whether their hypotheses about converging, diverging, or compensating operations are falsifiable (i.e., through measurement validation approaches). Further, we provide empirical examples along with access to open data sources available to investigators to test their hypotheses (see Table A3 in the online supplement).

SCD. SCD is a scientifically rigorous method for generating and testing hypotheses regarding anticipated discrepant evidence (Talbott & De Los Reyes, 2022). Literacy research provides a relevant example of this. As mentioned, in early reading, children acquire foundational literacy skills that allow them to build an orthographic lexicon that includes the spellings of words they already know and the spellings of words they learn through experience (Perfetti, 1992; Share, 1995). In many studies, researchers instantiate—and measure—foundational skills using a word recognition construct called by the IES "alphabetics." This construct includes pronunciation of real words and the ability to pronounce previously unknown words via decoding (e.g., Share, 1995). The alphabetics approach to determining the effects of instructional programs for word recognition takes a convergent approach and indexes both word pronunciation and decoding skills as part of the word recognition factor (e.g., Adlof et al., 2006; Høien-Tengesdal, 2010; Scarborough, 2002). This widely held view of word recognition is predicated on a converging operations interpretation of the links between word pronunciation and decoding—skills that converge to some extent but may differ from each other in domain-relevant ways (Hoover & Gough, 1990; Tunmer & Chapman, 2012).

Contrary to the converging operations interpretation, extensive data suggest that word pronunciation and decoding skills are distinguishable components of word recognition, suggesting that important information may be lost if the principle of converging operations governs interpretations of assessment data reflecting these two skill domains. In various theoretical accounts (Ouellette & Beers, 2010; Perfetti & Hart, 2002) and even neurobiological descriptions of word recognition (Kearns et al., 2019), word pronunciation and decoding play distinct roles in reading. For example, students with dyslexia frequently show different patterns of difficulty across these domains, typically with better performance on word pronunciation measures than on decoding measures (Frith, 1985). Some researchers have even used the distinction between decoding and word recognition to create reading phenotypes (e.g., phonological and surface dyslexia; Castles & Coltheart, 1993) that specifically distinguish readers based on their relative strengths and difficulties in word pronunciation and decoding.

Although researchers have traditionally assumed that the concept of converging operations provides the most accurate account of word recognition, word pronunciation and decoding skills are quite different. This distinction is particularly important when trying to understand unexpected reading difficulties of upper-elementary-age students. Examination of performance of word pronunciation and decoding measures often reveals the cause: These students have relatively good skills for reading high-frequency words—words that appear on many word pronunciation tests—but they have very weak decoding skills and thus no mechanism for adding new words to the lexicon. Students with dyslexia are likely to show this pattern, which can indicate the presence of a severe phonological processing deficit or the absence of instruction that focuses student attention on the connection between letters and sounds (Kearns & Whaley, 2019).

SCD methods provide researchers with the opportunity to test hypotheses about whether diverging operations may be present in assessment data germane to understanding word recognition. This hypothesis testing can be accomplished through the application of comparative designs, such as adapted alternating-treatment design (AATD), that allow an efficient comparison of effective instructional practices designed to address nonreversible behavior, such as reading skills (Wolery et al., 2018). AATD allows researchers to compare the effects of two interventions (e.g., word recognition and decoding) on separate behavior sets (e.g., a collection of discrete responses, such as high-frequency word lists; Wolery et al., 2018). Through targeted recruitment of study participants and detailed descriptions of their reading characteristics, hypotheses about interventions addressing anticipated discrepant outcomes can be tested within individuals who have dyslexia (Castles & Coltheart, 1993).

CONTEXT: Measurement validation paradigm. We previously mentioned that the CONTEXT validation paradigm is an empirical approach designed to guide the construction of validation studies that are capable of detecting domain-relevant data when findings converge (i.e., converging operations) but also when they meaningfully diverge (i.e., diverging operations). Yet, how might CONTEXT guide research practices when designing validation studies and interpreting their results? To address this question, we provide a brief description of this paradigm along with an empirical example; a larger description and expanded set of examples exist elsewhere (De Los Reyes et al., 2023).

As a validation paradigm, CONTEXT guides researchers to (a) pose conceptually grounded reasons for why discrepant estimates might contain domain-relevant information; (b) create falsifiable hypotheses to probe these reasons; (c) select data sources that, collectively, produce domain-relevant discrepant estimates; (d) select validity criteria that facilitate probing the presence of domain-relevant data in discrepant estimates; and (e) construct analytic models that allow researchers to test for the validity of indices of discrepant estimates while at the same time preserving the domain-relevant information contained in estimates of converging findings. Each of these features of CONTEXT is exemplified in recent work seeking to understand discrepant estimates as they manifest in assessments of adolescent social anxiety—assessments that we have long known commonly produce discrepant estimates (De Los Reyes et al., 2012). In this area of work, researchers have long questioned the veracity of adolescents' self-reports, with the idea that these discrepancies were the by-product of adolescents' motivations to provide socially desirable responses (see also De Los Reyes et al., 2015). In this respect, researchers commonly attributed discrepant estimates to measurement confounds, namely, a rater bias. However, researchers have also long known of the possibility that discrepant estimates of adolescent social anxiety could seemingly arise due to domain-relevant processes.

Consider the notion of situational specificity (Achenbach et al., 1987). If youth behave differently, depending on variations across contexts in the contingencies that "pull" for specific behaviors (e.g., anxiety), and informants vary in terms of which contexts and contingencies they are capable of observing (e.g., parent at home, teacher at school), then the discrepant estimates produced between informants' reports could contain domain-relevant data about how behavior changes within and across contexts. Only recently has this notion been applied to interpreting discrepant estimates in measures of adolescent social anxiety, and there was good reason all along to do so. Specifically, relative to youth earlier on in development, adolescents spend a greater amount of time outside of the home, and as such, adolescents who experience social anxiety often do so in contexts outside of parents' immediate presence, namely, within interactions with same-age, unfamiliar peers (Cannon et al., 2020). Recent work has involved developing controlled observation paradigms that seek to gather data about adolescents' reactions to these interactions with unfamiliar peers and in a way that facilitates gathering reports from untrained informants, namely, the peer confederates who, unlike parents, observe adolescents in these interactions and make reports about adolescents' social anxiety based on these interactions (Deros et al., 2018; Rezeppa et al., 2021). A particularly innovative element of this work stems from the additional gathering of ratings from trained research personnel, which serve as the validity criteria used to interpret data from the various informants (i.e., parents, adolescents, peer confederates). In this body of work, not only do discrepant estimates exist to a significant extent across these informants; these discrepancies also facilitate boosting the criterion-related validity of the informants' reports. In fact, recent work has involved constructing analytic models designed to capitalize on domain-relevant discrepant estimates (Charamut et al., 2022). These models produce scores that explain variance in domain-relevant validity criteria (i.e., ratings from trained, independent research personnel), over and above both each individual informant's report and the simple average of informants' reports (Makol et al., 2020). Taken together, recent work on understanding and interpreting discrepant estimates produced in assessments of adolescent social anxiety exemplify CONTEXT's core features.

Quality Indicators for EBA Tools and Processes

As we have seen, the OTM and CONTEXT can be used to guide measurement validation procedures that are tailored to the various purposes of assessment. As such, the OTM and CONTEXT are applied in concert with quality indicators for assessment outlined in Table A2. Here, we describe quality indicators for EBA, outlined in Table A2 and designed to guide measurement validation processes that may be used in tandem with the OTM and CONTEXT.

Central to the EBA process and measurement validation is the key indicator of construct validity. Construct validity, illustrated in Table A2, refers to the extent to which a measure assesses a given domain, trait, or characteristic (Kazdin, 2016); provides connections between theory and psychometric research (Borsboom et al., 2004); and illuminates the relationship between a given construct and the measures designed to assess it (Kazdin, 2016). As such, the measurement validation process begins with construct validity and requires the accumulation of evidence from diverse sources aligned with theory (Borsboom et al., 2004; Kazdin, 2016). The steps for developing a measure can be found in the psychometric characteristics for validity and reliability/precision outlined in Table A2 (Kazdin, 2016) and their associated rubric (De Los Reyes & Langer, 2018; Hunsley & Mash, 2007; Youngstrom et al. 2017).

In addition to the EBA quality indicators in Table A2 for validity and reliability, the table also includes guidance on norms, intervention sensitivity, clinical utility, and selecting outcome measures. Sources for the indicators include the *Standards for Educational and Psychological Testing* (AERA, 2014), *Common Guidelines for Education Research and Development* (U.S. Department of Education, 2013), SEER standards (U.S. Department of Education, 2022a), and the *WWC Procedures and Standards Handbook* (U.S. Department of Education, 2022b). The comprehensive EBA framework we propose necessitates use of these quality indicators. That is, measurement validation research guided by the OTM and relevant educational theory requires building upon, contextualizing, and applying established indicators to special education research.

The EBA standards in Table A2 are prescriptive. They guide researchers in identifying, developing, and evaluating scores taken from instrumentation (Jacobson & Svetina, 2019; Kazdin, 2019; Youngstrom et al., 2017). Although *validity* is associated with the interpretation of scores taken from a given instrument, *validation* is the primary method for understanding how that assessment works (AERA, 2014). The validation process must be substantive, conceptually grounded, and thus driven by factors beyond methodological considerations (Borsboom et al., 2004). Although various "types" of validity are referenced throughout the education and psychological literatures and reflected in the EBA quality indicators in Table A2 (i.e., content, construct, discriminative, predictive, face validity; validity generalization), these validity "types" may be more accurately viewed as research procedures for validation (Borsboom et al., 2004; Jacobson & Svetina, 2019). That is, validation refers to the testing and interpretation of measures for particular uses and is not a property of the measure itself (AERA, 2014; Jacobson & Svetina, 2019). Rather, the emphasis is on the validity of drawing inferences from scores taken from instruments (e.g., is it valid to interpret scores taken from measure X as indicating a given level of domain Y?). Researchers must apply these indicators when developing and selecting measures *and* engage in measurement validation to detect the key features of a construct captured by instrumentation (Kazdin, 2016). In this way, researchers can increase the likelihood of observing discrepant estimates that are best explained by diverging operations.

Not only can the measurement validation process be applied to interpreting scores taken from individual instruments; the process can and has been similarly applied to understanding patterns of data from diverse sources. This process involves researchers identifying criterion variables that function as independent measures of a given domain to serve as reference points for detecting the degree to which patterns observed across data sources reflect domain-relevant information (see De Los Reyes, Talbott, et al., 2022; De Los Reyes, Tyrell, et al., 2022; L. S. Fuchs et al., 2008; Lerner et al., 2017; Wakschlag et al., 2008). In fact, this is a key reason why the independence of validity criteria plays such a key role in the research practices that stem from the use of CONTEXT (see also De Los Reyes et al., 2023). Fortunately, special education researchers have multiple tools at their disposal to conduct these independent assessments for validation, samples of which we provide in the supplemental materials.

Researchers have multiple tools at their disposal to conduct independent assessments for validation, samples of which we provide in supplemental materials.

Empirical Contexts for the EBA

To advance EBA in special education research, we provide guidance within two central empirical contexts designed to identify and implement EBPs: (a) empirical research with a causal explanation and (b) implementation science research in special education. Figure A4 in the supplemental materials provides a graphical depiction of the EBA process within these two empirical contexts.

Empirical Research for Causal Explanation

The WWC (U.S. Department of Education, 2020) and the CEC (2014) have identified several research designs that, if well executed, result in causal inference: (a) randomized controlled trials (RCTs) and quasi-experimental designs, (b) SCD, and (c) regression discontinuity designs. Each of these designs represents an appropriate context for advancing EBA. That said, we focus on RCTs and quasi-experimental designs to help sharpen guidance from the field in the development and selection of outcome measures.

Clemens and D. Fuchs (2022) have outlined a measurement framework for the development and selection of outcome measures within a group design to assess the outcomes of an intervention (e.g., in reading comprehension). Different outcome measures will have different degrees of alignment (i.e., correlations) with the core components of a given intervention program. Although researcher-designed outcome measures may be more closely aligned with the researchers' intervention program than are commercial measures, these researcher-designed measures can be seen as appropriate for measuring a single, initial target skill that the program is designed to teach, followed by subsequent measurement tools in the framework. These subsequent tools, both researcher- made and commercial, are selected to measure additional target skills, as researchers plan to systematically assess near-transfer, mid-transfer, and far-transfer of skills taught during the intervention according to their theory of change and hypothesized intervention effects (Clemens & D. Fuchs, 2022). In the context of EBA, researchers must operationalize, clarify, and justify the use of both researcher-made and commercial measures in the context of their measurement framework, guided by a theory of change (Clemens & D. Fuchs, 2022; Youngstrom et al., 2017).

The OTM and CONTEXT can also serve to guide the development and selection of these measures, driven by findings from previous research. For example, one can identify outcome measures that meet quality indicators by examining correlations from prior meta-analyses, such as those in Table A1. These data help researchers anticipate the extent to which measures of a given construct are likely to converge, as may be the case in correlations between measures of reading comprehension and grammar knowledge in second-language learning (mean $r = .85$; Jeon & Yamashito, 2014), spelling and word reading (mean $r = .78$; Swanson et al., 2003), and curriculum-based measures of reading tasks with commercial standardized tests of reading (mean $r = .63$; Shin & McMaster, 2019). Conversely, researchers can anticipate many more examples (see Table A1) in which measures of a given construct are likely to diverge, as is the case in multi-informant measures of youth mental health (mean $r = .28$; De Los Reyes et al., 2015), in correlations between oral language skills in one's first language and these same skills in one's second language (mean $r = .16$, Melby-Lervag & Lervag, 2011), and in correlations between the use of cognitive strategies in learning with diverse academic achievement measures (mean $r = .11$; Dent & Koenka, 2016). Using prior research as a guide, we may also anticipate the presence of compensating operations, as may be the case with academic progress monitoring in the areas of reading and writing. That is, although prior guidance from research has been for practitioners to conduct curriculum-based measurement weekly, it now appears that there may be greater accuracy in measures conducted intermittently (Gesel & Lemons, 2020) and monthly (Hier et al., 2020). Future research needs to explore these compensating-operations scenarios, using preregistration as a mechanism for doing so. The OTM and CONTEXT may facilitate identifying and interpreting these measurement patterns to guide future research on what they might mean.

Implementation Science Research in Special Education

Implementation science is the study of methods to promote EBA as part of routine practice in special education; it means "using deliberate strategies in specific settings to adopt new interventions, integrate them effectively, and change practice patterns" (Lyon, 2016, p. 1). As a result, implementation research can take a variety of approaches, from a systems-oriented approach (e.g., multi-tiered systems of support; Odom et al., 2020) to approaches that are personalized and individualized to advance the academic, behavioral, and mental health goals for children and youth with disabilities (i.e., Bruhn et al., 2020; D. Fuchs et al., 2014; Hart, 2016; Kearns et al., 2021). We describe an explicit role for assessment in implementation science in Table A4 in the supplemental materials, with an overview of this process in Figure A4. This process begins with a needs assessment using multiple data sources, identified through attention to quality indicators in EBA tools and processes and guided by prior work within the OTM. This needs assessment is closely linked to goal setting and identifying EBPs for use within specific settings and includes ongoing engagement in assessment and intervention with local stakeholders (De Los Reyes, Talbott, et al., 2022).

A precise examination of the effective ingredients of EBPs, variously known as common elements or practice elements in child and adolescent mental health (e.g., Chorpita & Daleiden, 2009; McLeod et al., 2017), as instructional practices in academic interventions (e.g., Gersten et al., 2009); and as focused intervention practices for youth with ASD (Sam et al., 2020), suggests that ongoing monitoring of the effectiveness of these elements may facilitate their immediate tailoring to meet students' needs. Examples of outcomes to monitor include the effects of specific engagement strategies with youth from low-income families in mental health service delivery (Becker et al., 2018) and the effects of reading comprehension interventions for low-performing secondary students from diverse backgrounds, particularly those who are English learners (Vaughn et al., 2019). We know from prior research that advancing reading outcomes for English learners in the late elementary grades through high school can be particularly challenging (Vaughn & Wanzek, 2014). Solutions to these challenges will require greater attention to EBA, including those quality indicators we have outlined in Table A2. For example, the difficulty of identifying students from Spanish-speaking homes who struggle with reading comprehension is partly rooted in the domains identified for measurement (i.e., word reading vs. language comprehension; Mancilla-Martinez, 2020) as well as those measures selected to identify reading difficulties. In the case of the latter, developers of commercial, standardized measures have rarely reported the proportion of their norming samples who speak more than one language (Luk & Christodoulou, 2016). Likewise, we lack evidence-based guidance on how to attend to native language proficiency in reading assessment (Francis et al., 2019). Therefore, strategic investments in EBA must be advanced to address the needs of all youth in special education, including those who experience linguistic diversity.

Future Directions for Research

The persistent finding that 10% to 25% of the population of youth with disabilities and those at risk continue to experience poor academic, behavioral, and mental health outcomes—despite having participated in evidence-based interventions—creates a sense of urgency (D. Fuchs & L. S. Fuchs, 2019; Maggin et al., 2016). Such urgency is only magnified for youth from diverse backgrounds who have been underserved in high-quality programming due to historic racial and economic inequities (Harry & Klingner, 2014). In this section, we highlight future directions for EBA research on behalf of all learners with and at risk for disabilities.

Future Directions in Empirical Approaches Through Open Science

To further develop and advance EBA tools and processes, researchers can employ open data sources to advance measurement validation research. Recent advances in federal funding for the development of data repositories, such as the LDBase at Florida State University funded by the National Institutes of Health, help to move this work forward, as do those repositories with a 60-year history, such as the Inter-university Consortium for Political and Social Research (ICPSR). We include sample data sets from these repositories along with additional data sources in Table A3 to advance work in measurement validation and hypothesis development. This sample of open data sets includes multiple raters' reports of behavior, multiple measures of academic performance, and systematic direct observations. These types of data are embedded within longitudinal studies and studies using RCTs; they feature children and youth from diverse backgrounds. As such, these data can be used to develop and test hypotheses designed to advance empirical approaches within the OTM, such as measurement validation research.

Likewise, as mentioned, preregistration of research is an empirical tool researchers can use to facilitate the advancement of empirical approaches within the OTM and CONTEXT. By preregistering study hypotheses and study methods, particularly descriptions of measurement tools and hypotheses regarding how and when specific outcomes using these tools are expected to converge or diverge (e.g., Lombardi et al., 2023), researchers can significantly advance evidence-based assessment in the education sciences.

Future Directions in Empirical Research for Causal Explanation

In the context of advancing group design research in special education, we implore researchers to remember that the validity of a given instrument exists not in isolation (as a search for "common outcome measures" might indicate) but in *context*, which includes the many and diverse community and school contexts where students live and learn (AERA, 2014; Talbott & De Los Reyes, 2022). As we have seen, measurement validation may be viewed as an empirical means of testing a theory (Borsboom et al., 2004). It is thus incumbent on those who design, select, and employ EBA tools and processes to strengthen the technical processes of their measures, using quality indicators in Table A2 as their guide (Clemens & D. Fuchs, 2022; Kazdin, 2016).

For example, L. S. Fuchs and colleagues (2008) conducted an RCT using a mathematics word problem–solving intervention with third graders, employing a combination of researcher-made and commercially designed tools within a measurement framework. They measured proximal and distal effects of the intervention conducted over the course of the third grade year, even as they also conducted measurement validation using a dynamic assessment approach. This work provides an excellent model of the potential to integrate EBA and EBP. Although few researchers may have the resources to engage in such comprehensive work, our sample of open data sources can facilitate comparable research in measurement validation. Future directions in empirical research for causal explanation include addressing the consequences of assessment, particularly the influence of assessment on teaching and learning, which is uniquely suited to special education (Messick, 1995; L. S. Fuchs et al., 2008).

Future Directions in Implementation Science Research

Implementation science is designed to guide empirical approaches in special education research, including methods to study and promote teamwork, communication, and collaboration among leaders and practitioners in routine settings, such as education and health care (Odom et al., 2020; Talbott et al., 2021). As such, the impact of assessment should not be restricted to one type of measure or program of research (Kazdin, 2016). We urge funders and researchers to shift their focus away from a search for a single gold standard of assessment and toward an emphasis on hypothesis development and empirical testing to determine what causes change in student outcomes (Clemens & D. Fuchs, 2022; Hart, 2016). Such an approach is urgent for English learners with disabilities, for whom research in evidence-based assessment and intervention remain woefully understudied, despite the challenges these learners face compared with their English-proficient peers (Mancilla-Martinez, 2020). Empirical testing is essential to advance a continuum of measurement and yield a more precise understanding of mechanisms that produce change.

In Table A4, we have outlined stages of research in implementation science, identifying future directions for research, which include the need for (a) research in dynamic assessment (Grigorenko & Sternberg, 1998); (b) research to guide goal setting, particularly in behavioral and mental health research (Bruhn et al., 2020; De Los Reyes et al., 2022a); (c) research designed to identify and sustain EBA with various stakeholders, particularly in diverse and underserved communities (Baumann & Cabassa, 2020); (d) development of additional tools for progress monitoring of behavior (Chafouleas et al., 2021); (e) expansion of research on decision rules for the adaptation of interventions and their evidence-based components (McLeod et al., 2017); and (f) continued research on assessment within data-based individualization to meet students' complex needs (Kearns et al., 2021).

Conclusion

We have sought to advance EBA research in special education through the introduction of a conceptual and empirically tested model, the OTM, along with empirical methods for testing the OTM (i.e., via SCD and CONTEXT), a comprehensive set of quality indicators drawn from the education and psychological sciences, and application of these EBA approaches to two empirical contexts for research in EBPs. In applying the OTM to correlations from special education research in the diverse domains of language, reading, writing, mathematics, executive functioning, and mental health (see Table A1), we have aimed to ensure that researchers consider all three options in the context of relevant theory, regardless of the research domain: converging operations, diverging operations, and compensating operations. This point is crucial, given that the selection of any given paradigm charts a path for research, which includes the identification of theoretical models, the selection of data sources, and the approach to measurement validation (De Los Reyes et al., 2023).

We are reminded that assessment and intervention in EBPs are intimately connected. To that end, we appreciate Kurt Lewin's sage advice, "if you want truly to understand something, try to change it" (Tolman, 1996, p. 31), even as we recognize that this advice stops short. To truly understand that "something," including whether, how, for whom, and in what contexts interventions designed to change "something" actually work, we must accelerate investments in EBA. Only then will our students, particularly those who experience the greatest and most complex needs, experience the benefits of an appropriately challenging education.

References

Achenbach, T. M. (2011). Commentary: Definitely more than measurement error: But how should we understand and deal with informant discrepancies? *Journal of Clinical Child and Adolescent Psychology*, *40*(1), 80–86. https://doi.org/10.1080/15374416.2011.533416

Achenbach, T. M. (2020). Bottom-up and top-down paradigms for psychopathology: A half-century odyssey. *Annual Review of Clinical Psychology*, *16*(1), 1–24. https://doi.org/10.1146/annurev-clinpsy-071119-115831

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*(2), 213–232. https://doi.org/10.1037/0033-2909.101.2.213

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the Simple View of Reading include a fluency component? *Reading & Writing*, *19*, 933–958. https://doi.org/10.1007/s11145-006-9024-z

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.aera.net/Portals/38/1999%20Standards_revised.pdf

American Psychological Association. (2021). *APA guidelines on evidence-based psychological practice in health care*. https://www.apa.org/about/policy/psychological-practice-health-care.pdf

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475–493. https://doi.org/10.1037/a0032475

Baumann, A. A., & Cabassa, L. J. (2020). Reframing implementation science to address inequities in healthcare delivery. *BioMed Central Health Services Research*, *20*(190). https://doi.org/10.1186/s12913-020-4975-3

Becker, K. D., Boustani, M., Gellatly, R., & Chorpita, B. F. (2018). Forty years of engagement research in children mental health services: Multidimensional measurement and practice elements. *Journal of Clinical Child and Adolescent Psychology*, *47*(1), 1–23. https://doi.org/10.1080/15374416.2017.1326121

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., Fernandez, T., & Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, *22*(1), 5–19. https://doi.org/10.1016/j.cbpra.2014.02.002

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bruhn, A. L., Wehby, J. H., & Hasselbring, T. S. (2020). Data-based decision making for social behavior: Setting a research agenda. *Journal of Positive Behavior Interventions*, *22*(2), 116–126. https://doi.org/10.1177/1098300719876098

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cannon, C. J., Makol, B. A., Keeley, L. M., Qasmieh, N., Okuno, H., Racz, S. J., & De Los Reyes, A. (2020). A paradigm for understanding adolescent social anxiety with unfamiliar peers: Conceptual foundations and directions for future research. *Clinical Child and Family Psychology Review*, *23*(3), 338–364. https://doi.org/10.1007/s10567-020-00314-4

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition*, *47*(2), 149–180. https://doi.org/10.1006/jecp.1998.2482

Chafouleas, S. M., Johnson, A. H., Riley-Tillman, T. C., & Iovino, E. A. (2021). *School-based behavioral assessment: Informing prevention and intervention* (2nd ed.). Guilford.

Charamut, N. R., Racz, S. J., Wang, M., & De Los Reyes, A. (2022). Integrating multi-informant reports of youth mental health: A construct validation test of Kraemer and colleagues' (2003) satellite model. *Frontiers in Psychology*, *13*, 911629. https://doi.org/10.3389/fpsyg.2022.911629

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology*, *77*(3), 566–579. https://doi.org/10.1037/a0014565

Clemens, N. H., & Fuchs, D. (2022). Commercially developed tests of reading comprehension: Gold standard or fool's gold? *Reading Research Quarterly*, *57*(2), 385–397. https://doi.org/10.1002/rrq.415

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, *79*(3), 135–144. https://doi.org/10.1177/001440291307900201

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*(3), 365–383. https://doi.org/10.1177/001440290907500306

Council for Exceptional Children. (2014). Standards for evidence-based practices in special education. *Exceptional Children*, *80*, 504–511. https://doi.org/10.1177/0014402914531388

De Los Reyes, A. (2011). More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *40*(1), 1–9. https://doi.org/10.1080/15374416.2011.533405

De Los Reyes, A., & Aldao, A. (2015). Introduction to the special issue: Toward implementing physiological measures in clinical child and adolescent assessments. *Journal of Clinical Child & Adolescent Psychology*, *44*(2), 221–237. https://doi.org/10.1080/15374416.2014.891227

De Los Reyes, A., Aldao, A., Thomas, S. A., Daruwala, S. E., Swan, A. J., Van Wie, M., Goepel, K. A., & Lechner, W. (2012). Adolescent self-reports of social anxiety: Can they disagree with objective psychophysiological measures and still be valid? *Journal of Psychopathology and Behavioral Assessment*, *34*(3), 308–322. https://doi.org/10.1007/s10862-012-9289-2

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, *74*, 74–89. https://doi.org/10.1016/j.jsp.2019.05.005

De Los Reyes, A., Drabick, D. A. G., Makol, B. A., & Jakubovic, R. (2020). Introduction to the special section: The Research Domain Criteria's units of analysis and cross-unit correspondence in youth mental health research. *Journal of Clinical Child and Adolescent Psychology*, *49*(3), 279–296. https://doi.org/10.1080/15374416.2020.1738238

De Los Reyes, A., & Epkins, C. C. (2023). Introduction to the special issue: A dozen years of demonstrating that informant discrepancies are more than measurement error. Toward guidelines for integrating data from multi-informant assessments of youth mental health. *Journal of Clinical Child and Adolescent Psychology*, *52*(1), 1–18. https://doi.org/10.1080/15374416.2022.2158843

De Los Reyes, A., & Langer, D. A. (2018). Assessment and the *Journal of Clinical Child and Adolescent Psychology*'s evidence base updates series: Evaluating the tools for gathering evidence. *Journal of Clinical Child & Adolescent Psychology*, *47*(3), 357–365. https://doi.org/10.1080/15374416.2018.1458314

De Los Reyes, A., & Makol, B. A. (2022). Interpreting convergences and divergences in multi-informant, multi-method assessment. In J. Mihura (Ed.), *The Oxford handbook of personality and psychopathology assessment* (2nd ed.). Oxford University Press.

De Los Reyes, A., & Ohannessian, C. M. (2016). Introduction to the special issue: Discrepancies in adolescent-parent perceptions of the family and adolescent adjustment. *Journal of Youth and Adolescence*, *45*(10), 1957–1972. https://doi.org/10.1007/s10964-016-0533-z

De Los Reyes, A., Talbott, E., Power, T., Michel, J., Cook, C. R., Racz, S. J., & Fitzpatrick, O. (2022). The needs-to-goals gap: How informant discrepancies in youth mental health assessments impact service delivery. *Clinical Psychology Review*, *92*. https://doi.org/10.1016/j.cpr.2021.102114

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*(1), 123–149. https://doi.org/10.1146/annurev-clinpsy-050212-185617

De Los Reyes, A., Tyrell, F. A., Watts, A. L., & Asmundson, G. J. G. (2022). Conceptual, methodological, and measurement factors that disqualify use of measurement invariance techniques to detect informant discrepancies in youth mental health assessments. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.931296

De Los Reyes, A., Wang, M., Lerner, M. D., Makol, B. A., Fitzpatrick, O., & Weisz, J. R. (2023). The operations triad model and youth mental health assessments: Catalyzing a paradigm shift in measurement validation. *Journal of Clinical Child and Adolescent Psychology*, *52*(1), 19–54. https://doi.org/10.1080/15374416.2022.2111684

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Deros, D. E., Racz, S. J., Lipton, M. F., Augenstein, T. M., Karp, J. N., Keeley, L. M., & De Los Reyes, A. (2018). Multi-informant assessments of adolescent social anxiety: Adding clarity by leveraging reports from unfamiliar peer confederates. *Behavior Therapy*, *49*(1), 84–98. https://doi.org/10.1016/jbeeth.2017.05.001

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. (2012). Embracing not erasing contextual variability in children's behavior: Theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*(5), 558–574. https://doi.org/10.1111/j.1469-7610.2012.02537.x

Endrew F. v. Douglas County School District RE-1, 580 U. S. (2017).

Francis, D. J., Rojas, R., Gusewski, S., Santi, K. L., Khalaf, S., Hiebert, L., & Bunta, F. (2019). Speaking and reading in two languages: On the identification of reading and language disabilities in Spanish-speaking English learners. *New Directions for Child and Adolescent Development*, *166*, 15–41. https://doi.org/10.1002/cad.20306

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. C. Marshall & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). Routledge. https://doi.org/10.4324/9781315108346

Fuchs, D., Fuchs, L. S., & Vaughn, S. (2014). What is intensive instruction and why is it important? *TEACHING Exceptional Children*, *46*(4), 13–18. https://doi.org/10.1177/0040059914522966

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology*, *100*(4), 829–850. https://doi.org/10.1037/a0012657

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children*, *85*(2), 126–128. https://doi.org/10.1177/0014402918811924

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment*, *15*(4), 508–520. https://doi.org/10.1037/1040-3590.15.4.508

Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111. https://doi.org/10.3102/0034654313499616

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*(3), 149–159. https://doi.org/10.1037/h0042992

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202–1242. https://doi.org/10.3102/0034654309334431

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*(2), 149–164. https://doi.org/10.1177/001440290507100202

Gesel, S. A., & Lemons, C. J. (2020). Comparing schedules of progress monitoring using curriculum-based measurement in reading: A replication study. *Exceptional Children*, *87*(1), 92–112. https://doi.org/10.1177/0014402920924845

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75–111. https://doi.org/10.1037/0033-2909.124.1.75

Harry, B., & Klingner, J. (2014). *Why are so many minority students in special education? Understanding race and disability in schools* (2nd ed.). Teachers College Press.

Hart, S. A. (2016). Precision education initiative: Moving toward personalized education. *Mind, Brain, and Education*, *10*(4), 209–211. https://doi.org/10.1111/mbe.12109

Hier, B. O., January, S. A. A., & Van Norman, E. R. (2020). A comparison of CBM-WE scoring metrics and progress monitoring frequency among second-grade students. *School Psychology Review*, *49*(3), 306–320. https://doi.org/10.1080/2372966X.2020.1763758

Høien-Tengesdal, I. (2010). Is the simple view of reading too simple? *Scandinavian Journal of Educational Research*, *54*(5), 451–469. https://doi.org/10.1080/00313831.2010.508914

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160. https://doi.org/10.1007/BF00401799

Hosp, J., Hosp, M. K., Howell, K. W., & Allison, R. (2014). *The ABCs of curriculum-based evaluation*. Guilford Press.

Hosp, J., Howell, K. W., & Hosp, M. K. (2016). *The ABCs of CBM* (2nd ed.). Guilford Press.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*(1), 29–51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419

Individuals With Disabilities Education Act, 20 U.S.C. §§ 1400 *et seq*. (2004).

Jacobson, E., & Svetina, D. (2019). Prescribing structure for validation arguments: Elemental, structural, and ecological validity. *Applied Measurement in Education*, *32*(1), 43–59. https://doi.org/10.1080/08957347.2018.1544137

Jensen-Doss, A. (2011). Practice involves more than treatment: How can evidence-based assessment catch up to evidence-based treatment? *Clinical Psychology: Science and Practice*, *18*(2), 173–177. https://doi.org/10.1111/j.1468-2850.2011.01248.x

Jensen-Doss, A., & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child & Adolescent Psychology*, *39*(6), 885–896. https://doi.org/10.1080/15374416.2010.517169

Jeon, E. H., & Yamashito, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Johnson, A. H., & Cook, B. G. (2019). Preregistration in single-case design research. *Exceptional Children*, *86*(1), 95–112. https://doi.org/10.1177/0014402919868529

Kazdin, A. E. (2016). Selecting measures for research investigations. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 205–226). American Psychological Association.

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy*, *117*, 3–17. https://doi.org/10.1016/j.brat.2018.11.015

Kearns, D. M., Feinberg, N. J., & Anderson, L. J. (2021). Implementation of data-based decision-making: Linking research from the special series to practice. *Journal of Learning Disabilities*, *54*(5), 365–372. https://doi.org/10.1177/00222194211032403

Kearns, D. M., Hancock, R., Hoeft, F., Pugh, K., & Frost, S. (2019). The neurobiology of dyslexia. *TEACHING Exceptional Children*, *51*(3), 175–188. https://doi.org/10.1177/0040059918820051

Kearns, D. M., & Whaley, V. M. (2019). Helping students with dyslexia read long words: Using syllables and morphemes. *TEACHING Exceptional Children*, *51*(3), 212–225. https://doi.org/10.1177/0040059918810010

Ledford, J. R., Ayres, K. M., Lane, J. D., & Lam, M. F. (2015). Identifying issues and concerns with the use of interval-based systems in single case research using a pilot simulation study. *The Journal of Special Education*, *49*(2), 104–117. https://doi.org/10.1177/0022466915568975

Lerner, M. D., De Los Reyes, D., Drabick, D. A. G., Gerber, A. H., & Gadow, K. D. (2017). Informant discrepancy defines discrete, clinically useful autism spectrum disorder subgroups. *Journal of Child Psychology and Psychiatry*, *58*(7), 829–839. https://doi.org/10.1111/jcpp.12730

Lombardi, A. R., Rifenbark, G. G., & Taconet, A. (2023). Quality indicators of secondary data analyses in special education research: A preregistration guide. *Exceptional Children*. https://doi.org/10.1177/00144029221141029

Luk, G., & Christodoulou, J. A. (2016). Assessing and understanding the needs of dual-language learners. In N. Lesaux & S. Jones (Eds.), *The leading edge of early childhood education: Linking science to policy for a new generation of pre-K* (pp. 67–90). Harvard Education Press.

Lyon, A. R. (2016). *Implementation science and practice in the education sector*. Substance Abuse and Mental Health Services Administration. https://education.uw.edu/sites/default/files/Implementation%20Science%20Issue%20Brief%20072617.pdf

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*(3), 127–137. https://doi.org/10.1177/1063426616661498

Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating multiple informants' reports: How conceptual and measurement models may address long-standing problems in clinical decision-making. *Clinical Psychological Science*, *8*(6), 953–970. https://doi.org/10.1177/2167702620924439

Mancilla-Martinez, J. (2020). Understanding and supporting literacy development among English learners: A deep dive into the role of language comprehension. *AERA Open*, *6*(1). https://doi.org/10.1177/2332858420912198

Martel, M. M., Eng, A. G., Bansal, P. S., Smith, T. E., Elkins, A. R., & Goh, P. K. (2021). Multiple informant average integration of ADHD symptom ratings predictive of concurrent and longitudinal impairment. *Psychological Assessment*, *33*(5), 443–451. https://doi.org/10.1037/pas0000994

McLeod, B. D., Sutherland, K. S., Martinez, R. G., Conroy, M., Snyder, P., & Southam-Gerow, M. A. (2017). Identifying common practice elements to improve social, emotional, and behavioral outcomes of young children in early childhood classrooms. *Prevention Science*, *18*, 204–213. https://doi.org/10.1007/s11121-016-0703-y

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*(1), 114–135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*, 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Odom, S. L., Hall, L. J., & Suhrheinrich, J. (2020). Implementation science, behavior analysis, and supporting evidence-based practices for individuals with autism. *European Journal of Behavior Analysis*, *21*(1), 55–73. https://doi.org/10.1080/15021149.2019.1641952

Olino, T. M., Bufferd, S. J., Doughtery, L. R., Dyson, M. W., Carlson, G. A., & Klein, D. N. (2018). The development of latent dimensions of psychopathology across early childhood: Stability of dimensions and moderators of change. *Journal of Abnormal Child Psychology*, *46*, 1373–1383. https://doi.org/10.1007/s10802-018-0398-6

Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*, 189–208. https://doi.org/10.1007/s11145-008-9159-1

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman. (Eds.), *Reading acquisition* (pp. 145–174). Lawrence Erlbaum.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma. (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins.

Podsakoff, N. P., Podaskoff, P. M., Mackenzie, S. B., & Klinger, R. L. (2012). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology*, *98*(1), 99–113. https://doi.org/10.1037/a0029570

Rezeppa, T., Okuno, H., Qasmieh, N., Racz, S. J., Borelli, J. L., & De Los Reyes, A. (2021). Unfamiliar untrained observers' ratings of adolescent safety behaviors within social interactions with unfamiliar peer confederates. *Behavior Therapy*, *52*(3), 564–576. https://doi.org/10.1016/j.beth.2020.07.006

Sam, A. M., Cox, A. W., Savage, M. N., Waters, V., & Odom, S. L. (2020). Disseminating information on evidence-based practices for children and youth with autism spectrum disorder: AFIRM. *Journal of Autism and Developmental Disorders*, *50*, 1931–1940. https://doi.org/10.1007/s10803-019-03945-x

Scarborough, H. S. (2002). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson. (Eds.), *Handbook of early literacy research* (Vol. 1, pp. 97–110). Guilford Press.

Schneider, M. (2020). *Making common measures more common* [Blog post]. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/director/remarks/5-05-2020.asp

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151−218. https://doi.org/10.1016/0010-0277(94)00645-2

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in *Educational and Psychological Measurement*. In B. D. Zumbo & E. K. H. Chan. (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91–112). Springer.

Shin, J., & McMaster, K. L. (2019). Relations between CBM and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, *73*, 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2019). *Evidence-based medicine: How to practice and teach EBM* (5th ed.). Elsevier.

Talbott, E., & De Los Reyes, A. (2022). Making sense of multiple data sources: Using single case research design for behavioral decision making. In T. W. Farmer, E. Talbott, K. L. McMaster, D. Lee, & T. C. Aceves. (Eds.), *Handbook of special education research: Theory, methods, and developmental processes (Vol. 1, pp.* 231–244). Routledge.

Swanson, H. L., D. M., G., Necoechea, D. M., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A Meta-analysis of the correlation evidence *Review of Educational Research*, 73(4), 407–440. http://dx.doi.org/10.3102/00346543073004407

Talbott, E., De Los Reyes, A., Power, T., Michel, J., & Racz, S. J. (2021). A team-based collaborative care model for youth with attention deficit hyperactivity disorder in education and health care settings. *Journal of Emotional and Behavioral Disorders*, *29*(1), 24–33. https://doi.org/10.1177/1063426620949987

Tamm, L., Loren, R. E. A., Peugh, J., & Ciesielski, H. A. (2021). The association of executive functioning with academic, behavior, and social performance ratings in children with ADHD. *Journal of Learning Disabilities*, *54*(2), 124–138. https://doi.org/10.1177/002221942061338338

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence based practice. *Exceptional Children*, *71*(2), 181–194. https://doi.org/10.1177/001440290507100204

Tolman, C. W. (Ed.). (1996). *Problems of theoretical psychology* (Vol. 6). Captus Press.

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*(5), 453–466. https://doi.org/10.1177/0022219411432685

U.S. Department of Education. (2013). *Institute of Education Sciences and the National Science Foundation: Common guidelines for education research and development*. https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

U.S. Department of Education. (2020). *What Works Clearinghouse standards handbook, Version 4.1.* https://ies.ed.gov/ncee/wwc/handbooks

U.S. Department of Education. (2022a). *Standards for excellence in education research*. https://ies.ed.gov/seer/

U.S. Department of Education. (2022b). *What Works Clearinghouse procedures and standards handbook, Version 5.0.* https://ies.ed.-gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Vaughn, S., Roberts, G., Capin, P., Miciak, J., Cho, E., & Fletcher, J. M. (2019). How initial word reading and language skills affect reading comprehension outcomes for students with reading difficulties. *Exceptional Children*, *85*(2), 180–196. https://doi.org/10.1177/0014402918782618

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research and Practice*, *29*(2), 46–53. https://doi.org/10.1111/ldrp.12031

von der Embse, N., & De Los Reyes, A. (2023). Advancing equity in access to school mental health through multiple informant decision-making. *Journal of School Psychology*. In press.

Wakschlag, L. S., Briggs-Gowan, M. J., Hill, C., Danis, B., Leventhal, B. L., Keenan, K., Egger, H. L., Cicchetti, D., Burns, J., & Carter, A. S. (2008). Observational assessment of preschool disruptive behavior, Part II: Validity of the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS). *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(6), 632–641. https://doi.org/10.1097/CHI.0b013e31816c5bdb

Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2022). How robust is the p-factor? Using multitrait-multimethod modeling to inform the meaning of general factors of psychopathology in youth. *Clinical Psychological Science*, *10*(4), 640–661. https://doi.org/10.1177/21677026211055170

White, O. R. (1986). Precision teaching-precision learning. *Exceptional Children*, *52*(6), 522–534.

Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparative designs. In J. R. Ledford & D. L. Gast. (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 283–316). Routledge.

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M. L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*, *24*(4), 331–363. https://doi.org/10.1111/cpsp.12207

Authors' Note

Efforts by the second and fifth authors were supported by a grant from the Institute of Education Sciences (R324A180032).

ORCID iDs

Elizabeth Talbott https://orcid.org/0000-0002-9163-3414

Devin M. Kearns https://orcid.org/0000-0001-9703-0932

Supplemental Material

The supplemental material is available in the online version of the article.

Open-Science Badge

For publishing their data, Talbott et al. received badge for open data. The public content may be retrieved from http://doi.org/10.17605/OSF.IO/H2AJK.

Manuscript received March 2022; accepted April 2023.