Contemp School Psychol (2014) 18:1–12
DOI 10.1007/s40688-013-0009-z

Applied Empiricism: Ensuring the Validity of Causal Response to Intervention Decisions

Stephen P. Kilgus · Melissa A. Collier-Meek · Austin H. Johnson · Rose Jaffery

Published online: 7 January 2014

Abstract School personnel make a variety of decisions within multitiered problem-solving frameworks, including the decision to assign a student to group-based support, to design an individualized support plan, or classify a student as eligible for special education. Each decision is founded upon a judgment regarding whether the student has responded to intervention. These and other conclusions are inherently causal, thus requiring that educators carefully consider the internal, construct, and conclusion validity of each decision to ensure its defensibility. Researchers have identified multiple variables that are likely to moderate these validities, including the integrity with which interventions are implemented, the psychometric adequacy of progress-monitoring tools, the extent to which interventions and supports are matched to a student's needs, and the approach to single-case research design. We therefore review each of these variables in the interest of assisting practitioners to design acceptable and valid multitiered frameworks of prevention and service delivery.

Keywords Response to intervention · Research methods · Validity

School-based multitiered frameworks, including response to intervention (RTI), have gained prominence as general service delivery models, emphasizing risk minimization through prevention, and early identification and intervention (Vaughn and Fuchs 2003). Recently, Kratochwill et al. (2012) reviewed the role of science within school psychological practice, illustrating the relationships between the RTI, scientist-practitioner, and evidence-based practice movements. The authors specified two main roles filled by school psychology scientist-practitioners. The first pertains to the consumption of scientific evidence. RTI is founded upon the use of evidence-based practices; schools are required to use interventions supported by scientific evidence, which serves as a preliminary indicator that such strategies are appropriate and likely to promote student outcomes. Judgment as to whether evidence is both scientific and sufficient to support applied use is a primary function of the scientist-practitioner (Kratochwill et al. 2012). The second role is the generation of scientific evidence. Training in research methods and data analysis permits the scientist-practitioner to engage in *practice-based research* for the purpose of collecting evidence of the effectiveness of interventions and supports (Sidman 2011). Such research calls for the use of sophisticated methods and analyses, which together support the validity of inferences regarding intervention effectiveness and increase the likelihood of the evidence's contribution to the peer-reviewed literature. Many have called for the proliferation of such research, as it stands to both enhance the dissemination of evidence-based practices and determine how practices may be adapted to fit local contexts and enhance effectiveness (Kratochwill et al. 2012).

Although a focus on school psychologists as generators of evidence is necessary and crucial, we see potential for expansion to the definition of the scientist-practitioner evidence-generation responsibilities described by Kratochwill et al. (2012). Despite its obvious benefits, the collection of evidence worthy of contributing to the scientific literature is frequently infeasible given the constraints of the applied setting Kazdin et al. (1986). This is not to say that use of research methods is not expected. Be it suitable for publication or not, evidence is necessary to support local RTI-related professional judgments and causal inferences regarding whether manipulation of an

S. P. Kilgus (✉)
East Carolina University, Greenville, NC, USA
e-mail: kilguss@ecu.edu

M. A. Collier-Meek · A. H. Johnson
University of Connecticut, Mansfield, CT, USA

R. Jaffery
EASTCONN, Hampton, CT, USA

independent variable (e.g., an intervention) resulted in a corresponding change in a dependent outcome variable (e.g., oral reading fluency; Riley-Tillman and Burns 2009). This less intense, albeit important, form of local research has been differentiated from practice-based research and described as *research translation* (Sidman 2011).

The methods and analyses used within the research translation process are less rigorous than their practice-based research counterparts. For instance, research translation is not likely to include randomization or extensive controls for extraneous variables (Sidman 2011). Yet, the process may still yield evidence supportive of multiple important functions. First, local research translation evidence may be used to determine whether an intervention yielded intended effects. Although an intervention may be hypothesized to work for a student given its support in the literature, prior local effectiveness, and fit to the student's needs, its effectiveness must be demonstrated and cannot be assumed (Riley-Tillman and Burns 2009). Use of experimental methods and collection of technically adequate evidence may permit such demonstrations, as they promote the various validity types that influence local decisional accuracy, including internal, construct, and conclusion validity (see Table 1 for a review of each of these types; Shadish et al. 2002). They also facilitate the problem-solving process through which each student's progress is monitored toward timely application of effective supports (Sidman 2011). Second, research evidence can serve a protective function, as it ensures each student is afforded due process protections (Noell and Gansle 2006). According to IDEA 2004, to classify a child with a learning disability, it must be shown that his or her underachievement is not the result of limited appropriate instruction. Within the RTI model, a school must therefore demonstrate that the lack of response driving a change in supports or placement was, despite use of interventions, appropriately matched to each student's needs (Riley-Tillman and Burns 2009). Technically adequate assessment tools and experimental controls must therefore be used to ensure interventions were appropriate and not responsible for the student's lack of response.

In summary, it can be argued that it is necessary to support RTI practices through basic research given its potential to not only promote the fairness and ethicality of decisions but also increase the effectiveness and timeliness of services. Stoner and Green (1992) described this local research as fulfilling the promise of the scientist-practitioner model by creating an “experimenting society,” wherein applied decisions are founded upon empiricism. Scientist-practitioners may consider multiple research elements suggested to influence the validity of causal inferences (Kratochwill et al. 2012). These include the use of (a) assessment to inform selection of appropriate interventions, (b) treatment integrity assessment, (c) single-case research designs and data analytic procedures, and (d) defensible, flexible, repeatable, and efficient progress-monitoring measures. The purpose of this paper is to review each of these elements, examine how each influences one or more validity types, and briefly describe practices that promote the validity of RTI decisions. It is hoped that this information will be useful to school psychologists working to promote the ethicality and effectiveness of their services.

Intervention Match

Within an RTI model, interventions should correspond to two broader construct categories to ensure the credibility and appropriateness of student-related decisions. First, each intervention should be *evidence-based.* The construct validity of such an inference is supported by scientific evidence indicative of the strategy's effectiveness. Second, each intervention should be *matched* to a student's demonstrated difficulties.

Table 1 Definitions of validity types related to the decisional accuracy of RTI-related causal inferences

Validity type	Definition	Factors that promote validity
Internal	Confidence with which one may conclude that changes in an independent variable (e.g., explicit timing) caused changes in one or more dependent variables (e.g., math computation fluency)	Use of experimental and quasi-experimental research designs, controls for extraneous variables, collection and interpretation of progress-monitoring data and collection of treatment integrity data
Construct	Appropriateness of generalizations from observable methods and procedures (e.g., use of a timer, instructions to work quickly and accurately for 1-min segments, and provision of feedback) to broader intervention categories (e.g., explicit timing, math intervention, and appropriately matched intervention).	Collection of data regarding match between a student's needs and selected supports and collection of treatment integrity data
Conclusion	Degree to which one may reasonably derive inferences regarding the existence and extent of relationships between variables in consideration of specific data and analyses	Appropriate use and interpretation of analyses of data that meet relevant assumptions. For instance, the use of baseline control techniques when visually analyzing naturally trending data (e.g., oral reading fluency) or the use of baseline standard deviation when calculating standardized mean differences and homogeneity of variance across phases cannot be assumed

Intervention match may be considered a multifaceted concept. In determining whether an intervention is a match, one must collects data to determine whether these are both (a) sufficiently intense and (b) functionally appropriate (Riley-Tillman and Burns 2009).

Intensity The question of intensity pertains to whether the applied supports are of sufficient strength to yield an intended effect of eliminating the discrepancy between the student's current and expected functioning (Yeaton and Sechrest 1981). Information regarding the extent of this discrepancy may be collected using one of two ways. First, within the academic domain, the degree of a student's difficulties is frequently examined through screening via curriculum-based measurement. Many schools have begun to conduct triannual benchmarking of academic functioning, utilizing the data to determine students' level of risk for learning difficulty (e.g., low, some, and high). The extent of risk may then correspond to interventions of varying intensity, with those at greater risk receiving more intense supports. A similar approach has been proposed within the emotional and behavioral domain through use of various methods, including selected screening via behavior rating scales or universal screening via multiple gating procedures or universal rating methods (Lane et al. 2012). Second, the appropriateness of intervention intensity may be evaluated within a problem-solving approach, wherein progress-monitoring data are collected and reviewed following intervention application. Data suggestive of an intervention's effectiveness indicate that the strategy is appropriately matched to the student's needs. If data indicate insufficient improvement, there may be a need for increased dosage or an alternate intervention. (For more information regarding progress monitoring, please see below.)

It should be noted that although ensuring an intervention is appropriately intense given a student's needs is important, there is still a fundamental assumption that intensity mismatch will occur. The existence of multiple tiers of service delivery, each corresponding to an increasingly intensive level of support, implies this inevitability. As such, the identification of interventions of appropriate intensity may not be feasible to support all lower-stakes decisions (e.g., tier 2 intervention selection). Yet, there is a need to ensure an attempt has been made to identify an appropriately intense intervention within the overall RTI process, as this would promote the construct validity of higher-stakes decisions (e.g., special education eligibility).

Function The second dimension along which an intervention may be matched to a student's needs is through its alignment with the function of the academic or behavioral problem. Within the academic domain, the cause of student difficulties may be considered in terms of reasonable functional hypotheses (e.g., insufficient motivation, prompting and feedback, and modeling; Daly et al. 1997). Research has supported the use of brief experimental analysis (BEA) as a means to test each of these functional hypotheses (Daly et al. 1997). The goal of the BEA procedure is to identify one or more interventions that are best matched to the student's needs and therefore most likely to be effective. An educator begins by forming hypotheses regarding the nature of the student's problem. Interventions associated with each hypothesis are then applied in a rapid and alternating fashion within an alternating treatment design (see below for more information regarding this approach). The intervention to which the student responds best is thought to be associated with the true function of the student's problem and should therefore be chosen for prolonged implementation. Daly et al. (1997) demonstrated the tenability of the approach, and subsequent studies have supported its use across multiple academic domains (e.g., Burns et al. 2009).

Within the behavioral domain, a fundamental assumption is that all behavior is purposeful and therefore maintained by the consequences that follow it (e.g., adult attention; Cooper et al. 2007). Researchers have called for educators to conduct functional behavior assessments (FBAs) to identify which consequences have maintained problem behavior and to use this information in the design and selection of functionally relevant interventions. This recommendation is founded in research that has shown function-based interventions to be more effective than their non-function-based counterparts (Filter and Horner 2009). Some have identified tier 3 as the level at which functional information should be considered. However, a line of research supports the utility of considering functional data at tier 2, as function has shown to moderate the effectiveness of multiple commonly targeted interventions (McIntosh et al. 2009). As it would be highly impractical to conduct full FBAs for all tier 2 students, it is recommended that educators instead only complete one or more efficient functional assessment methods that are typically incorporated into FBAs, including functional interviews, rating scales, and checklists. Although the indirect nature of these procedures is likely to yield relatively less valid decisions, the low-stakes nature of tier 2 decisions and the need for functional information may support their use.

In summary, the need to ensure intervention match should not be understated, as it is necessary if RTI is to be a substantive model of service delivery (Noell and Gansle 2006). The promise of RTI lies in the timely application of supports to prevent student problems. Little benefit is gained from providing a student with supports without some indication that they are likely to be appropriate. The provision of potentially ineffective supports runs the risk of wasting time during which problems might worsen and collecting evidence that may lead to RTI-related inferences lacking construct validity (e.g., “The student did not respond to *appropriate* interventions”). It could therefore be argued that time would be better spent engaging in additional assessment and problem identification prior to implementation and using this information to make more informed decisions regarding intervention selection (Fuchs et.al. 2003). Although intervention success is never guaranteed, the demonstration of match increases confidence in the likelihood of effectiveness and the validity of inferences (Riley-Tillman and Burns 2009).

Single-Case Design

Whereas the immediate focus of traditional group design research is to yield results that are generalizable to a greater population of interest, research utilizing single-case design (SCD) is devoted to the study of the sample itself, investigating whether educators can have confidence that an independent variable was responsible for a change in a dependent variable. In other words, does a functional relationship exist between the independent variable (e.g., group contingency intervention) and dependent variable (e.g., student disruptive behavior)? Each time the dependent variable systematically changes in response to the application or removal of the independent variable (e.g., from a baseline phase to an intervention phase), an effect is demonstrated. There is also a related increase in confidence regarding a functional relationship between the two variables and the internal validity of inferences regarding this relationship.

Design Elements The internal validity of inferences regarding functional relationships is related to multiple design elements, each of which should be considered relative to the research question of interest and its associated stakes when implementing SCDs. First, because repeated measurements of the dependent variable are necessary to determine if the intervention was effective, educators need to determine how many data points are sufficient. Recommendations from The What Works Clearinghouse's (WWC) technical report on SCD (Kratochwill et al. 2010) suggest the collection of at least three data points, but preferably five, in each phase of a design. Doing so permits an initial understanding of trend and more accurate estimation of variability and level in performance within each phase. It is thus recommended for the results of a study to meet WWC standards for evidence. Second, educators should consider how many demonstrations of an effect are necessary to support conclusions regarding a functional relationship. As with data points, the “magic number” for demonstrations of an effect continues to be three (Horner et al. 2005). If three opportunities for a demonstration of an effect are provided and the study otherwise possesses sufficient controls for extraneous variables, then one should have confidence in the internal validity of causal conclusions drawn from that study.

Design Types In an RTI framework, the question of “how much confidence is needed?” may be relative to the tier at which an intervention is being implemented. As described by Riley-Tillman and Burns (2009), practices at tiers 1, 2, and 3 may be equated to and supplemented by SCD designs. Tier 1 is similar to a simple B design in that supports (e.g., positive reinforcement of school-wide behavioral expectations) are consistently in place. Each measurement, be it a test score or behavior rating, occurs while an intervention is underway. At tier 2, what was occurring at tier 1 may now be considered a baseline. This baseline is then compared with what occurs following implementation of the tier 2 intervention, establishing an AB design. Tier 3 supports are the most intensive, and the decisions regarding their effectiveness tend to be higher stakes. As such, there is now a need for at least three opportunities to demonstrate an effect to accurately determine whether the intervention was effective. Of course, providing the opportunity for three demonstrations of an effect is no simple task; that being said, steps may be taken to limit the time and resources associated with the procedure. Three particular SCD types provide an opportunity for three demonstrations of an effect and may be feasible at scale in applied settings. Which of these is the most appropriate within any particular instance is tied to the nature of the case and dependent variable of interest.

The first and most well-known design is the ABAB/reversal design. This design provides three opportunities for a demonstration of an effect by transitioning between baseline and intervention for three times: The intervention is applied, then removed, and then applied again. The reversal design has a history of use within both academic and behavioral research and is considered an intuitive and powerful means to demonstrate the relationship between a treatment and a change in an outcome. Yet, although easily understood and highly common within educational research, the use of reversal designs is discouraged under certain circumstances. For instance, reversal designs are considered inappropriate when ethical considerations preclude ceasing an intervention, such as with a student who is engaging in behavior that is potentially harmful to his or her self or others or if a positive response is observed upon initial implementation and stakeholders are wary of removing an effective intervention. Reversal designs are also considered ill-suited when the target of intervention is irreversible, as they might lead to effectiveness-related inferences lacking in conclusion validity.

For example, assume that the effectiveness of a spelling intervention is to be evaluated through a reversal design. A change in spelling from baseline conditions to the intervention supports an initial conclusion of the intervention's success. However, a reversal to baseline conditions is not associated with a return to baseline spelling levels, as the student did not “lose” the spelling skill he or she acquired during the intervention phase. Based upon interpretive guidelines for reversal designs, the initial conclusion would be inappropriately rejected given the absence of the verification of baseline predictions necessary to establish a functional relationship between the intervention and spelling ability (Cooper et al. 2007). Therefore, though reversal design can be a flexible and feasible SCD, when the dependent variable of interest is likely irreversible or a reversal to baseline might be unethical, it is recommended that educators consider using alternate SCDs.

The second common design that provides for the opportunity for three demonstrations of an effect is the multiple baseline design. Multiple baseline designs are particularly relevant when the dependent variable of interest is irreversible or requiring a return to baseline condition will be infeasible or inappropriate (Hammond and Gast 2010). In this design, the intervention is not removed in order to provide the opportunity for a demonstration of an effect but rather implemented in a staggered fashion across at least three “conditions” of interest for the case. If a demonstration of an effect is observed in each condition under each intervention implementation, then the interventionist may confidently infer a functional relationship. Some possible conditions across which one could replicate include students, settings, and behaviors. As seen here, some of the most potent arguments for using multiple baseline result from its flexibility; for instance, the multiple baseline design offers the opportunity to add cases during a “study.” As long as there is a sufficient degree of overlap in baseline and intervention data collection across conditions, inferences may be derived regarding functional relationships. This allowance for the addition of conditions over time makes multiple baseline particularly well suited for use at tier 2, where students are continuously added and removed from existing targeted interventions over time.

The third common design, alternating treatments, is particularly well suited for simultaneously testing for a functional relationship with multiple interventions. It is frequently used to quickly identify which of two or more interventions is most effective in remediating a student's difficulties. Similar to the reversal design, alternating treatments design requires that the dependent variable in question be reversible. In contrast to reversal design and similar to multiple baseline design, alternating treatments design does not require a reversal to baseline conditions. In alternating treatments, a baseline phase occurs as usual, but in the second phase, treatments are rapidly alternated across sessions in a randomized fashion (e.g., treatment A during session 1, treatment B during sessions 2 and 3). Crucial to the integrity of this design is identifying conditions over which these interventions will be implemented and balancing them accordingly (Cooper et al. 2007). For example, if treatment A was always administered in the morning and treatment B in the evening, the effect of time-of-day could not be disentangled from that of the treatments.

Data Analysis We raise a brief note regarding the analysis of SCD data. At this time, visual analysis remains the gold standard for the interpretation of single-case research. Yet, research has noted difficulties associated with the derivation of valid visual analyses of SCD data, indicating the need for advanced training and understanding of appropriate interpretive guidelines, such as controlling for a trend in baseline data (Knapp 1983; Mercer and Sterling 2012). Partly in recognition of these findings, recent research has targeted the development of methods for the statistical analysis of graphed data. Such analyses are of particular interest in the case of single-case evidence syntheses and potentially in educator quantification of student response to intervention. Effect sizes commonly used in SCD research include percentage of nonoverlapping data (PND; Scruggs, Mastropieri, and Casto 1987), standardized mean difference (Busk and Serlin 1992), and improvement rate difference (IRD; Parker et al. 2009). However, little agreement has been reached on how well these effect sizes function (Kratochwill et al. 2010). Research indicates that certain effect size statistics are particularly prone to bias under certain conditions. For instance, PND may drastically underestimate the effect of an intervention in the presence of variable or trending baseline data (Campbell 2004). It is therefore recommended that effect sizes be considered as supplemental to visual analysis rather than as a stand-alone analytic tool. Sole reliance upon potentially biased effect sizes may reduce the conclusion validity of causal inferences, resulting in inappropriate conclusions regarding whether and to what extent an intervention was effective. For more information on visual analysis and effect sizes, see Riley-Tillman and Burns (2009).

Taken together, although somewhat time and resource intensive, there is a clear justification to utilize SCD in applied settings to support valid causal inferences. Given the relatively low stakes of many tier 1 and 2 decisions, the use of less rigorous designs (e.g., simple B and AB) may be appropriate. Yet, as the stakes of decisions increase, so too does the need for experimental rigor, and educators may subsequently consider utilizing more sophisticated SCD types that will allow for additional demonstrations of an effect. Although somewhat cumbersome, certain designs are particularly amenable to the RTI structure. Furthermore, their use may provide information useful in making more informed and effective service delivery decisions (Sidman 2011).

Treatment Integrity

Treatment integrity data indicate the extent to which an intervention is implemented as planned (Gresham 1989; Sanetti and Kratochwill 2009). Unfortunately, adequate levels of treatment integrity levels are often assumed, rather than assessed (Cochrane and Laux 2008). In a survey of 806 nationally certified school psychologists, respondents reported that treatment integrity data were collected during only 11.3 % of one-to-one consultation and 1.9 % of team consultation, even though 97.6 % of those surveyed endorsed the assessment of treatment integrity as important (Cochrane and Laux 2008). These rates are concerning, as research indicates that higher levels of treatment integrity are associated with better student outcomes (Biggs et al. 2008), and most teachers demonstrate variable and low levels of treatment integrity within only 2 weeks following intervention training (Noell et al. 2005). As such, failure to assess the independent variable has serious implications for student outcomes and data-based decision making.

Implementation data are particularly important within RTI, where decisions regarding supports and services are based on inferences regarding the rate and magnitude of student response to intervention. Treatment integrity data support the construct and internal validity of causal inferences regarding intervention effectiveness (Peterson et al. 1982). Absence of such information limits an educator's ability to determine whether the intervention was implemented as intended, if it is responsible for a change in the student and whether the intended intervention should be modified or replaced given limited student response. For instance, data indicative of a lack of student response may support development of a more intensive intervention (see the intervention match method described above). However, if treatment integrity data indicate that only 60 % of the critical components of an intervention are regularly implemented, a more appropriate strategy would likely be to bolster intervention implementation.

Assessing Treatment Integrity There are few prescribed treatment integrity assessment measures. Therefore, for most cases, assessment measures specific to the intervention and context must be developed (Schulte et al. 2009). In the absence of specific research-derived rules for developing these measures, five steps to guide the design of treatment integrity measures have been proposed (for further detail see Sanetti et.al. 2011). First, after an evidence-based intervention is selected, the specific steps should be delineated and described in operational, behavioral terms (Gresham 1989). Second, in consideration of these steps, a feasible and appropriate assessment method should be chosen (i.e., direct observation, permanent product review, and self-report; Sanetti et al. 2011). These methods vary in their (a) time intensity, (b) obtrusiveness, and (c) objectivity. The designer should consider the implementer and the intervention and perhaps use a combination of techniques. For example, a self-monitoring intervention may result in permanent products that could be reviewed to assess implementation; however, observation may provide a less biased estimate of a teacher's delivery of contingent praise. Third, in consideration of the steps and the assessment method, one should determine how delivery of each intervention step will be rated. Step delivery may be rated using dichotomous (yes/no) or Likert-type scales (e.g., 1–5), among other methods. Fourth, the designer should consider how often treatment integrity data will be assessed and evaluated. Like the other guidelines described here, there is no clear rule. However, similar to student outcome data, one may consider the intensity of the intervention and what decisions will be made based on the data (Sanetti et al. 2011). Treatment integrity assessment might be completed more often and with more rigorous methods when complex, intensive interventions are implemented or when high-stakes decisions might be made based on the data. Last, one should regularly evaluate treatment integrity data.

Reporting Treatment Integrity Treatment integrity data are not regularly included in the school psychology literature, potentially due to the lack of unified recommendations around reporting these data (Sanetti and Kratochwill 2009). Until consensus around reporting emerges, educators may consider including treatment integrity mean and range across sessions (i.e., overall treatment integrity and session integrity), as well as the treatment integrity mean and range of specific intervention components (i.e., component integrity). In addition, educators might consider reporting treatment integrity data related to dimensions of treatment integrity outside of adherence (e.g., quality, participant responsiveness) and describing if promotion strategies, such as performance feedback, were used.

Progress Monitoring

Collection and evaluation of progress-monitoring data are essential for determining whether an intervention should be maintained, modified, or discontinued (Brown-Chidsey et al. 2009). School psychologists can serve a vital role in assisting school problem-solving teams to determine how to measure change in target outcomes. Emphasis should be placed on consideration of the psychometric defensibility of potential progress-monitoring tools, as this may influence the internal validity of causal claims. Consideration of unreliable and invalid data may lead to RTI-related decisions lacking internal validity, given that changes in student performance attributed to introduction of an intervention may be explained by measurement error or a change in some unspecified and inadvertently measured variable.

Psychometric Defensibility A search of any academic search engine will reveal many progress-monitoring tools across multiple domains. According to the National Center on Response to Intervention (NCRTI; www.rti4success.org), research supports conducting academic progress monitoring via both general outcome measures (e.g., Dynamic Indicators of Basic Early Literacy Skills) and mastery measures (e.g., Accelerated Math). Within the behavioral domain, commonly used progress-monitoring methods and procedures include systematic direct observation (e.g., momentary time sampling), short-form behavior rating scales (e.g., change-sensitive brief behavior rating scales), Direct Behavior Rating (DBR; e.g., single-item scales; Gresham 2010), and goal attainment scaling (Ruble et.al. 2012). When deciding which of these specific academic or behavioral tools are to be used in monitoring a student's response to intervention, educators must determine whether each potential candidate possesses multiple psychometric properties outlined by the NCRTI and Urbina (2004), among others, as necessary for supporting the applied use of a progress-monitoring tool.

First, each tool must yield scores that evidence acceptable *reliability.* Of the numerous reliability types, several are relevant to academic progress-monitoring tools, including test-retest, internal consistency, and inter-rater. Availability of alternate forms that produce reliable data should also be considered, as it affords frequent data collection. Effective progress monitoring uses formative assessment or several short and quick evaluations of student performance over time. Such data need to be collected frequently (e.g., daily and weekly) to examine interventions and make adjustments based on student needs. The more intense the problem and the more high stakes the decision, the more frequently data should be collected and interpreted. Thus, the availability of parallel forms that are equivalent in level of difficulty and sufficient for frequent administration may be a necessary component of academic progress-monitoring tools.

Whether each of these reliability types is pertinent to behavioral progress-monitoring tools depends on the specific tool in question and circumstances associated with data collection. For instance, test-retest may be expected of certain measures, including change-sensitive brief behavior rating scales, given their occasional pertinence to more stable trait-like attributes (e.g., social competence). It may be less expected of DBR data, as the state-like variables targeted by the tool (e.g., disruption) are likely to vary over time in response to changing environmental contingencies. It is therefore recommended that educators carefully consider the progress-monitoring tool of interest and determine which reliability evidence is necessary to support its applied use. Relevant technical manuals and peer-reviewed research should then be considered in search of pertinent evidence.

Second, each tool should demonstrate acceptable *validity,* with evidence supporting both the measure's correspondence to variables it is intended to measure as well as its use in decision making. Multiple validity types are relevant to both academic and behavioral progress-monitoring tools, including content, criterion-related, and construct validity. According to Messick (1995), each of these validity types may actually be considered to fall under the broader umbrella of construct validity, which provides the evidential basis for score use. Construct validity evidence is also part of a broader unified validity framework, which takes into consideration a measure's relevance/utility, value implications, and social consequences (Messick 1995). Although each of these latter facets is far less frequently examined in the literature, existing reviews set a precedent for their consideration in providing support for progress-monitoring tools (e.g., Good and Jefferson 1989).

Third, *sensitivity* to student improvement toward goals is another feature of defensible progress-monitoring tools. The tool must be sensitive enough to measure incremental changes to demonstrate intervention effects. To ensure this, data obtained by the tool must be functionally related to the target outcome and matched to the skills being taught in the intervention. Effective progress-monitoring tools also specify *rates of improvement* expected of students making sufficient growth with regard to benchmarks. Such information is vital during periodic performance evaluations throughout the year and for setting student goals. Although evidence regarding sensitivity to change is available in both the academic and behavioral progress-monitoring literature, there is an absence of evidence regarding expected rates of improvement within the behavior literature. This supports the need for idiographic decision making within the behavioral realm given the absence of norm- or criterion-referenced information.

Additional Considerations Once educators have selected a psychometrically defensible progress-monitoring tool, they must ensure data will be collected consistently. Chafouleas et al. (2007) outlined additional factors to consider that may improve the reliability and validity of data and therefore strengthen the internal validity of causal decisions made based on data obtained from the tool. First, individualized academic or behavioral targets of the intervention must be operationally defined and measurable. Definitions must be objective and detailed so that both data collectors and interpreters can recognize and understand each target. Second, educators must provide opportunities for data collectors to be trained for mastery on using the selected progress monitor. Data collectors should also be provided booster-training sessions to maintain high training standards and minimize artifacts (e.g., observer drift and recency effects). Finally, data should be collected in a predetermined, controlled, and periodic fashion.

Case Example

Intervention Match The following case example illustrates how each of the aforementioned recommendations might be incorporated to increase the internal, construct, and conclusion validity of RTI-related inferences. Mrs. Ramirez, a third grade teacher at Northeast Elementary School, approached the school psychologist with concerns regarding three students in her classroom. Mrs. Ramirez indicated that the students had been engaging in a range of problem behaviors. Suspecting that these behaviors were indicative of risk for externalizing problems, the school psychologist asked Mrs. Ramirez to complete the *Student Risk Screening Scale* (Drummond 1994), a brief seven-item screener, for each of the students. SRSS ratings indicated that each student was at moderate risk for externalizing problems, suggesting the necessity of tier 2 behavioral supports. As a follow-up, the school psychologist met with Mrs. Ramirez to complete the *Functional Assessment Checklist: Teachers and Staff* (FACTS; March et al. 2000). Results of the brief interview revealed two main problem behaviors common to each of the students: disruption and off-task behavior. FACTS findings further indicated that these problem behaviors were most common during morning seatwork and math lecture and were primarily maintained by adult attention (student 1), escape from academic tasks (student 2), and peer attention (student 3).

Based upon the results of the problem identification assessment, it was determined that Check In/Check Out (CICO) was likely to be an appropriately intense intervention for all students. Typical CICO procedures were applied to the student whose behaviors were maintained by adult attention, as CICO has been found to be most effective for these individuals (McIntosh et al. 2009). Additional minor function-based modifications were made to CICO procedures for the remaining two students to increase the relevance of the intervention to their behavioral function (Campbell and Anderson 2008; Turtura et.al. 2013).

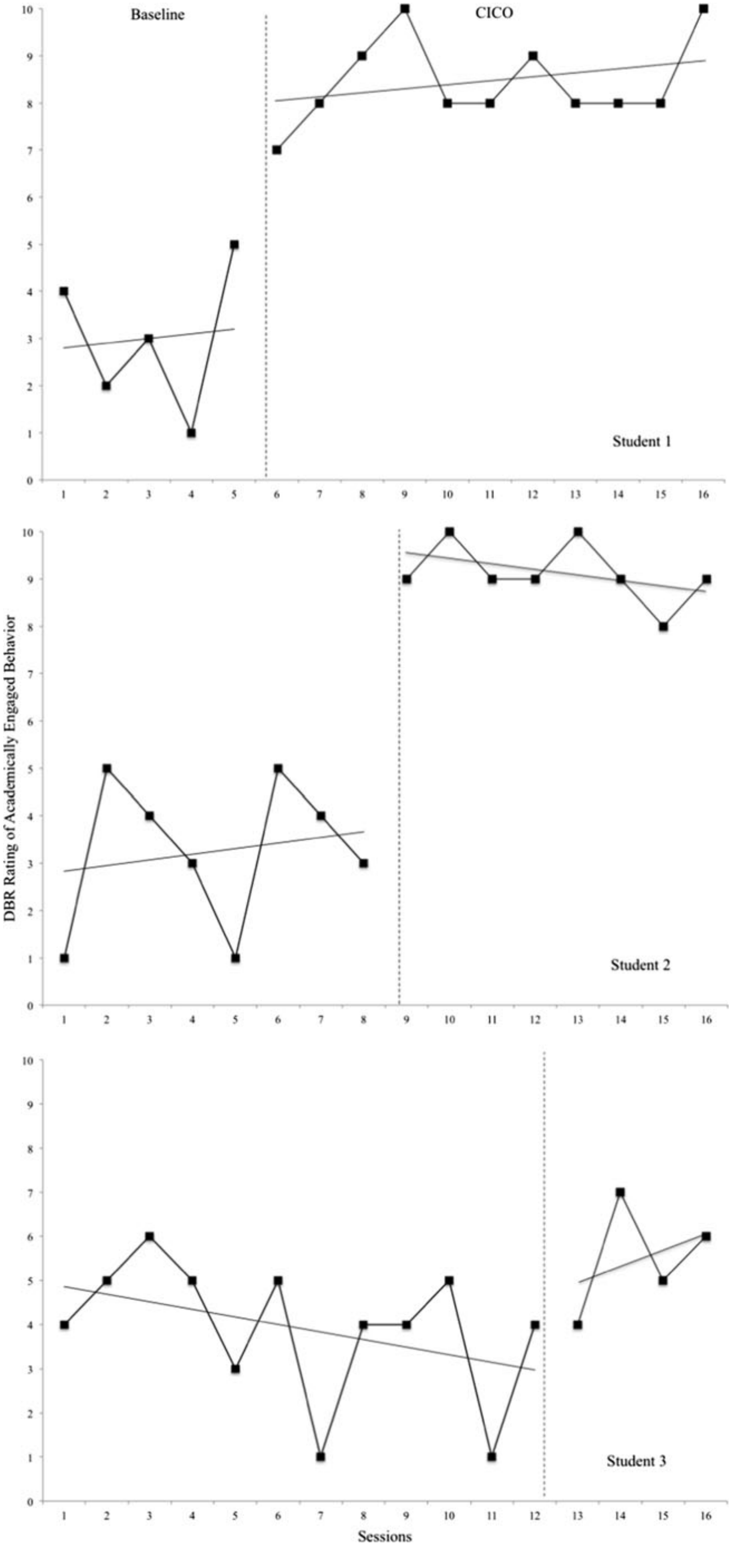
Treatment Integrity The school psychologist chose to monitor Mrs. Ramirez's treatment integrity via two methods. First, whether Mrs. Ramirez regularly rated each student's behavior, and appropriately delivered reinforcing stimuli was evaluated through a review of permanent products (i.e., completed CICO daily progress reports). Second, the school psychologist conducted one direct observation of Mrs. Ramirez's CICO implementation per week. Of specific interest was whether the nature and content of Mrs. Ramirez's feedback to the students were in accordance with the intervention protocol. The school psychologist also established a treatment integrity criterion of 80 %. If treatment integrity assessments indicated that Mrs. Ramirez had correctly implemented fewer than 80 % of intervention steps for three consecutive days, the school psychologist consulted with Mrs. Ramirez to increase the fidelity with which CICO was implemented.

Progress Monitoring and SCD Next, the school psychologist and Mrs. Ramirez collaboratively selected DBR single-item scales as a progress-monitoring tool. This decision was based upon a review of the literature, which was indicative of support for DBR criterion-related validity, generalizability and dependability, and sensitivity to change (Chafouleas 2011). Using DBRs, Mrs. Ramirez rated each student's disruptive behavior and academically engaged behavior twice per day within the two settings considered to be most problematic for the three students. Data collection began at the same time for all students as part of the baseline phase of a multiple baseline design. Multiple baseline was considered most appropriate in evaluating CICO effectiveness given Mrs. Ramirez's expressed desire to not revert to baseline conditions once intervention began. Students were staggered into the intervention, with the order of introduction based upon the stability of baseline data.

Data Analysis DBR data were evaluated through two methods. First, data were visually analyzed in terms of the level, trend, and variability of data. CICO effectiveness was evaluated through the analysis of changes in these characteristics across baseline and intervention phases. Conclusions regarding the presence of experimental control were based upon a review of the similarity across students in terms of (a) baseline characteristics, in an evaluation of whether initial baseline predictions were subsequently verified, and (b) intervention characteristics, in an evaluation of whether treatment effects were replicated across the students. Second, IRD effect sizes were calculated to supplement visual analysis in the comparison of baseline and intervention performance for each student. IRD was chosen in recognition that relative to the related PND statistic, IRD is less biased in the presence of outliers and atypical baseline performance (Parker et al. 2009).

See Fig. 1 for a graphical summary of example findings as they pertain to academically engaged behavior. A visual analysis of the findings indicated that CICO was effective for students 1 and 2, with academic engagement moving from variable and low to consistent and high following the introduction of CICO. CICO was not as effective for student 3, whose academic engagement remained variable and low. IRD statistics corroborated these findings, with large effect sizes for students 1 and 2 reflecting minimal data overlap across the phases (IRD=1.00 for each) and a small effect size for student 3 connoting the high degree of overlap and minimal change (IRD=0.50). Based upon student 3's response, the school psychologist reengaged in problem identification assessment to verify the original functional hypothesis. Direct observations supported the assumption that the problem behavior was maintained by peer attention, thereby indicating that the student's nonresponse likely resulted from the use of an intervention that was not sufficiently intensive. Student 3 was therefore referred for assessment and intervention at tier 3.

Fig. 1 Example data demonstrating effect of Check In/ Check Out on academic engagement



Application to School Psychologists

For RTI to function as an effective framework for service delivery, the decisions to provide students with particular levels of support must be well founded. More specifically, RTI decisions should have adequate validity. As school psychologists hold unique training, expertise, and responsibilities, they are particularly qualified to help ensure the appropriateness of RTI decisions and associated supports for students (Kratochwill et al. 2012; National Association of School Psychologists 2010). In their multifaceted role as scientist-practitioners, school psychologists can work to incorporate appropriate assessment of intervention match, treatment integrity assessment, single-case research design, and psychometrically defensible progress-monitoring methods into the RTI framework at their schools.

To do so, school psychologists can participate as active members of problem-solving teams to advocate for these elements to be incorporated into individual student plans as well as the system-wide RTI framework; additional resources to help guide school psychologists through these processes are listed in Table 2. For example, during problem-solving team meetings, school psychologists may encourage utilizing SCDs that are appropriate to the students' level of risk and tier of support. In addition, the appropriateness of an intervention for a particular student can be assessed by engaging in problem identification procedures and through a review of progress-monitoring data. Before an intervention is implemented, a school psychologist may advocate for the use of psychometrically defensible progress-monitoring methods and treatment integrity assessment. Throughout implementation, he or she may support ongoing data collection to evaluate the effectiveness of an intervention.

This article is a primer to several critical components of RTI. As this framework is relatively new, many practicing school psychologists may not feel conversant in the topics or be able to fluently apply them to their practice. In this case, it may be appropriate to seek out advanced training in these areas, whether it be the identification of appropriate progress-monitoring measures, the implementation of advanced SCDs, or the defensible use of statistic or visual analytic techniques or evidence-based problem identification procedures (e.g., FBA). In this way, school psychologists may incorporate the elements described here to promote the validity of RTI decisions within their setting. As RTI is, at heart, little more than the foundation for making decisions about how to best promote student success in schools, these methodologies may provide the bricks and mortar for engaging in actionable work with students in schools.

Table 2 Elements to promote the validity of RTI decisions: key components and additional resources

Research element	Resources
Intervention match	
○ Interventions should be *appropriately intense* to address student needs, which can be assessed through regular benchmarking of student risk level or a problem-solving approach ○ Interventions should be *appropriate to the function of the academic or behavior issues*, which can be assessed through brief experimental analysis or functional behavior assessment	Crone and Horner 2003; Riley-Tillman and Burns 2009
Treatment integrity	

○ Treatment integrity data indicate the extent to which an intervention was implemented as planned
○ These data are necessary to accurately evaluate student outcome data and make informed data-based decisions
	Researchers without Borders (http://www.researcherswithoutborders.org);
	Heartland Area Education Agency (http://www.aea11.k12.ia.us/idm/checkists.html);
	Sanetti et al. 2011

Single-case design	
○ Single-case designs inform whether a functional relationship exists between the independent variable (e.g., an intervention) and dependent variable (e.g., student oral reading fluency and academic engagement) ○ Appropriate designs (and related confidence) may vary at different tiers, with increasing tiers associated with more rigorous evaluations of effect	Alberto and Troutman 2009; Riley-Tillman and Burns 2009
Progress monitoring	

○ Progress-monitoring measures should be psychometrically defensible, in that each measure should have acceptable reliability, validity, and sensitivity
○ Other considerations include ensuring behavioral targets are sufficiently operationalized, proper training for data collectors is provided, and data are collected regularly
	Chafouleas et al. 2007;
	Intervention Central (http://www.interventioncentral.org);
	National Center on Response to Intervention (NCRTI; www.rti4success.org)

References

Alberto, P. A., & Troutman, A. C. (2009). *Applied behavior analysis for teachers* (8th ed.). Upper Saddle River: Pearson.

Biggs, B. K., Vernberg, E. M., Twemlow, S. W., Fonagy, P., & Dill, E. J. (2008). Teacher adherence and its relation to teacher attitudes and student outcomes in an elementary school-based violence prevention program. *School Psychology Review, 37*, 533–549.

Brown-Chidsey, R., Bronaugh, L., & McGraw, K. (2009). *RTI in the classroom: guidelines and recipes for success.* New York: Guilford.

Burns, M. K., Ganuza, Z. M., & London, R. M. (2009). Brief experimental analysis of written letter formation: single-case demonstration. *Journal of Behavioral Education, 18*, 20–34.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: new directions for psychology and education* (pp. 187–212). Hillsdale: Erlbaum.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234–246.

Campbell, A., & Anderson, C. M. (2008). Enhancing effects of check-in/check-out with function-based support. *Behavioral Disorders, 33*, 233–245.

Chafouleas, S. M. (2011). Direct behavior rating: a review of the issues and research in its development. *Education and Treatment of Children, 34*, 575–591.

Chafouleas, S. M., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: informing intervention and instruction.* New York: Guilford.

Cochrane, W. S., & Laux, J. M. (2008). A survey investigating school psychologists' measurement of treatment integrity in school-based interventions and their beliefs about its importance. *Psychology in the Schools, 45*, 499–507.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River: Pearson.

Crone, D. A., & Horner, R. H. (2003). *Building positive behavior support systems in schools: functional behavioral assessment.* New York: Guilford.

Daly, E. J., Witt, J. C., Martens, B. K., & Dool, E. J. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review, 26*, 554–574.

Drummond, T. (1994). *The student risk screening scale.* Grants Pass: Josephine County Mental Health Program.

Filter, K. J., & Horner, R. H. (2009). Function-based academic interventions for problem behavior. *Education and Treatment of Children, 32*, 1–19.

Fuchs, D. M., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*, 157–171.

Good, R. H., & Jefferson, G. (1989). Contemporary perspectives on curriculum-based measurement validity. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61–88). New York: Guilford.

Gresham, F. M. (2010). Data-based decision making for students' social behavior. *Journal of Evidence-Based Practices for Schools, 11*, 149–168.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*, 37–50.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983–2007. *Education and Training in Autism and Developmental Disabilities, 45*, 187–202.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Kazdin, A. E., Kratochwill, T. R., & Vanden Bos, G. (1986). Beyond clinical trials: generalizing from research to practice. *Professional Psychology: Research and Practice, 3*, 391–398.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155–164.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). Single-case designs technical documentation. Resource document. What Works Clearinghouse.: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf. Accessed 18 Aug 2012

Kratochwill, T. R., Hoagwood, K. E., Kazak, A. E., Weisz, J. R., Hood, K., Vargas, L. A., et al. (2012). Practice-based evidence for children and adolescents: advancing the research agenda in schools. *School Psychology Review, 41*, 215–235.

Lane, K. L., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction: from preschool to high school.* New York: Guilford.

March, R. E., Horner, R. H., Lewis-Palmer, T., Brown, D., Crone, D., Todd, A. W., et al. (2000). *Functional Assessment Checklist: Teachers and Staff (FACTS).* Eugene, OR: Educational and Community Supports.

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions, 11*, 82–93.

Mercer, S. H., & Sterling, H. E. (2012). The impact of baseline trend control on visual analysis of single-case data. *Journal of School Psychology, 50*, 403–419.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

National Association of School Psychologists (2010). Model for comprehensive and integrated school psychology services. *Communique, 39*(4), 1–6. Resource document.: http://nasponline.org/standards/2010standards/2_PracticeModel.pdf. Accessed 18 Aug 2012

Noell, G. H., & Gansle, K. A. (2006). Assuring the form has substance: treatment plan implementation as the foundation of assessing response to intervention. *Assessment for Effective Intervention, 32*, 32–39.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., et al. (2005). Treatment implementation following behavioral consultation in schools: a comparison of three follow-up strategies. *School Psychology Review, 34*, 87–106.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*, 135–150.

Peterson, L., Homer, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*, 477–492.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: single-case design for measuring response to intervention.* New York: Guilford.

Ruble, L., McGrew, J. H., & Toland, M. D. (2012). Goal attainment scaling as an outcome measure in randomized controlled trials of psychosocial interventions in autism. *Journal of Autism and Developmental Disorders, 42*, 1974–1983.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: introduction to the special series. *School Psychology Review, 38*, 445–459.

Sanetti, L. M. H., Fallon, L. M., & Collier-Meek, M. A. (2011). Treatment integrity assessment and intervention by school-based personnel: practical applications based on a preliminary study. *School Psychology Forum, 5*, 87–102.

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38*, 460–475.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: methodology and validation. *Remedial and Special Education, 8*, 24–33.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Sidman, M. (2011). Can an understanding of basic research facilitate the effectiveness of practitioners? Reflections and personal perspectives. *Journal of Applied Behavior Analysis, 44*, 973–991.

Stoner, G., & Green, S. K. (1992). Reconsidering the scientist-practitioner model for school psychology practice. *School Psychology Review, 21*, 155–166.

Turtura, J. E., Anderson, C. M., & Boyd, R. J. (2013). Addressing task avoidance in middle school students: academic behavior check-in/check-out. *Journal of Positive Behavior Interventions, 42*((6)), 1–9. doi:10.1177/1098300713484063.

Urbina, S. (2004). *Essentials of psychological testing.* Hoboken: Wiley.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: the promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137–146.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156–167.

Stephen P. Kilgus, Ph.D., NCSP is an assistant professor in the school psychology program in the Thomas Harriot College of Arts and Sciences at East Carolina University. He received his Ph.D. in school psychology from the University of Connecticut. His primary research interests include emotional and behavioral assessment, targeted behavior interventions and supports, and the statistical analysis of screening and diagnostic tests.

Melissa A. Collier-Meek, PhD is a postdoctoral fellow serving as Project Manager for Project PRIME at the Center for Behavioral Education and Research (CBER) at University of Connecticut. In addition, she works as a behavior consultant at EASTCONN, where she supports school teams in implementing evidence-based practices. Her research interests include treatment integrity assessment and implementation in schools and other settings.

Austin H. Johnson, MA is a pre-doctoral intern at EASTCONN, a regional educational service center in Northeastern Connecticut. As a fifth-year student in school psychology at the University of Connecticut, his dissertation research focuses on sources of variance in ratings derived from systematic direct observation procedures.

Rose Jaffery, Ph.D., NCSP is a nationally certified school psychologist and behavioral consultant in northeastern Connecticut. Her primary areas of research interest include behavioral assessment, multi-tiered systems-level change, collaboration across home, school, and community settings, and evidence-based strategies for improving the quality of life for children with autism.