Fidelity Measurement in Consultation: Psychometric Issues and Preliminary Examination

Article *in* School psychology review · December 2009

CITATIONS
60

READS
551

5 authors, including:

Susan M. Sheridan
University of Nebraska at Lincoln
226 PUBLICATIONS **4,666** CITATIONS
SEE PROFILE

Greg Welch
University of Nebraska at Lincoln
14 PUBLICATIONS **327** CITATIONS
SEE PROFILE

Kyongboon Kwon
University of Wisconsin - Milwaukee
20 PUBLICATIONS **437** CITATIONS
SEE PROFILE

School Psychology Review,
2009, Volume 38, No. 4, pp. 476–495

Fidelity Measurement in Consultation: Psychometric Issues and Preliminary Examination

Susan M. Sheridan, Michelle Swanger-Gagné, Greg W. Welch, Kyongboon Kwon, and S. Andrew Garbacz
Nebraska Center for Research on Children, Youth, Families and Schools
University of Nebraska—Lincoln

Abstract. Consultation researchers have long recognized the importance of assessing fidelity of intervention implementation, including the fidelity with which both consultation procedures and behavioral intervention plans are delivered. However, despite decades of discussion about the importance of assessing for fidelity of implementation in intervention delivery, the empirical foundation lags far behind in systematic efforts to incorporate reliable, valid, and conceptually meaningful fidelity measurement into its procedures. The methods used to capture elements of implementation are often incomplete, imprecise, and of questionable reliability. Among the methods commonly used to assess intervention fidelity in consultation (self-report, permanent products, direct observation), there exists little to no research documenting their psychometric adequacy. This article explores issues surrounding the assessment of fidelity in consultation research, including its rationale and role in consultation and intervention science. Methods for conceptualizing and assessing fidelity, psychometric issues, and research needs are identified. The results of a descriptive, exploratory study tapping the reliability of fidelity assessment measures within the context of a large-scale efficacy trial are presented, with a call for rigorous research to advance the consultation field.

In this era of increased demands for accountability and heightened standards for effective interventions, researchers must be concerned with both the availability of treatments to bolster student performance and the evaluation of their effects. There is an increasing push toward using evidence-based practices, and concomitantly, more consumers are expecting highly effective treatment plans (Drake et al., 2001; Frese, Stanley, Kress, & Vogel-Scibilia, 2001). At the same time, researchers are under pressure to demonstrate that their interventions contribute to a body of treatments or services that can be expected to

Preparation of this paper was supported in part by a grant awarded to the first author by the U. S. Department of Education, Institute of Education Sciences (Grant R305F050284). The opinions stated are those of the authors and should not be construed as representing those of the funding agency.

Correspondence regarding this article should be addressed to Susan M. Sheridan, Department of Educational Psychology, 239 Teachers College Hall, University of Nebraska—Lincoln, Lincoln, NE 68588-0345; E-mail: ssheridan2@unl.edu

produce important, desired effects (Mowbray, Holter, Teague, & Bybee, 2003). To adequately and reliably test the efficacy of interventions or treatment programs (Dane & Schneider, 1998), it is necessary to understand if intervention implementation is actually occurring as designed. Variations in implementation fidelity have been shown to contribute to programming outcomes (Durlak, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003; Zvoch, Letourneau, & Parker, 2007); therefore, determination of the effect of consultation-based intervention is not possible without knowledge of factors associated with implementation fidelity. In this article, we explore issues surrounding the assessment of fidelity in consultation research, including its rationale and role in consultation and intervention science. Methods for conceptualizing and assessing fidelity, psychometric issues, and research needs are identified. Likewise, we present the results of an exploratory, descriptive study that taps the reliability and validity of fidelity assessment measures within the context of a large-scale efficacy trial, noting limitations and further research needs.

Careful measurement of intervention implementation fidelity in consultation-based interventions is essential for several reasons. In the most basic sense, our ability to infer that an intervention is effective (i.e., an outcome is a function of the treatment variable, or is internally valid) requires knowledge about its implementation. For researchers and practitioners to make valid, causal inferences about an intervention's effect (i.e., have confidence in the relationship between a treatment and behavioral outcome), we must first know what constitutes the independent variable (i.e., treatment). Indeed, treatment effects can only be described relative to the degree to which the treatment was delivered and received (Cordray & Pion, 2006). As fidelity of an otherwise efficacious intervention deteriorates, our ability to draw any inferences about treatment effects diminishes. Thus, failure to measure intervention implementation fidelity may lead to an incorrect conclusion that an intervention protocol is *in*effective when in fact it *is* effective (or conversely, as effective when it is not).

In a practical sense, lack of attention to fidelity may lead to implementation of the "wrong" treatment. This is both a theoretical and empirical problem, as it results in an evaluation of the effects of an intervention as *described,* rather than *delivered,* yielding unreliable results with little to no bearing on actual intervention effects (a "Type III error"; Dobson & Cook, 1980). Furthermore, insufficient assessment of implementation fidelity may mask the reality that an intervention was applied in a nonsystematic, variable fashion, and hinder replication under similar conditions using identical procedures (Rossi & Freeman, 1985).

Considerations of Fidelity in Consultation-Based Research

Consultation as a service delivery model is widely used in school psychology and related disciplines to address behavioral, academic, and social-emotional concerns experienced by children. In practice, behavioral consultation is a structured problem solving process with ample evidence of efficacy and acceptability (Guli, 2005; Sheridan, Clarke, & Burt, 2008). An extension of behavioral consultation, conjoint behavioral consultation (CBC; Sheridan, Kratochwill, & Bergan, 1996; Sheridan & Kratochwill, 2008), is procedurally operationalized through a series of well-defined stages (including problem identification, problem analysis, plan development, plan evaluation), wherein consultees develop and implement coherent, coordinated intervention plans across home and school settings. Like more traditional forms of behavioral consultation (e.g., Bergan & Kratochwill, 1990), the model is oriented toward helping identified students achieve identified goals, using data collected throughout the process to monitor response to the specified intervention, and supporting consultees in developing competencies to address similar problems in the future. In addition, CBC is concerned with strengthening family–school partnerships, promoting continuity between home and

school, and enhancing parent engagement in educational collaboration.

CBC as an intervention for addressing identified student concerns has received extensive empirical support, primarily through experimentally controlled, small-*n* designs (for reviews, see Guli, 2005; Sheridan et al., 2008; and Sheridan, Eagle, Cowan, & Mickelson, 2001). However, relative to other processes required in its implementation (e.g., target behavior specification, determination of data collection procedures, evaluation of treatment effects), the model provides little attention or guidance toward treatment fidelity (Erchul & Schulte, 1996; Wilkinson, 2006). Thus, the lack of reliable and valid methods available for assessing fidelity within the behavioral consultation framework is not surprising.

The nature and effects of consultation service delivery are inherently challenged by several fundamental realities that complicate fidelity assessments. First, consultation interventions are delivered in naturalistic field settings, and rarely are they standardized with formal manuals or protocols. Second, in "real-life" contexts, treatment plans are often comprised of multiple components delivered in a conditional manner (e.g., given behavior *x*, deliver *y*), requiring knowledge and determination of both antecedent and sequential events. Third, in consultation, consultees (e.g., parents and teachers) are responsible for implementation of intervention plans, rather than highly trained behavioral therapists. Consultees are often untrained in the procedures comprising the planned intervention, and sometimes unmotivated or poorly supported within the larger context of the treatment setting (e.g., school system). Fourth, target interventions are nested in a complex system of other environmental events and realities, such as classroom routines, school supports or staffing issues, agency/district policies, community norms, family cultural values, and a host of other ecological variables. Any one or more of these may influence what is experienced at the individual child level (Bronfenbrenner, 1979). Finally, interventions are designed and implemented in highly individualized, idiosyncratic means within consultation. The logic of behavioral and functional assessments warrants individualized plans constructed to address specific conditions and functions. Thus, reliability and validity of assessment tools are often unique to one or a small number of cases, and traditional forms of determining reliability and validity are inapplicable.

Operationalization of the construct of fidelity within the context of consultation is complex. Although definitions abound, the realities of consultation services contribute to variations in core aspects of a treatment (e.g., what it involves, how it is to be delivered, who is responsible and involved, and how much is enough). Fit of the intervention (match to target behavior, theoretical justification, and acceptability by consultees and clients), implementation processes (including plan specification and qualifications of the treatment agent), congruity between a treatment and its expected outcomes (based on anticipated, research-supported effects under varying levels of strength), and parents' and teachers' uptake (receptivity to or engagement with the intervention) all influence fidelity considerations (Cordray & Pion, 2006). These realities within consultation complicate a researcher's ability to determine the degree to which a treatment is delivered with integrity.

A second issue in the consideration of treatment fidelity within consultation concerns the fact that the consultation process itself is a separate aspect of, and potential source of variance in, the intervention sequence. Consultation may be considered a "two-tiered" intervention (i.e., independent variable), with fidelity issues obvious at each of the tiers. The first tier represents the consultation model that is being implemented (e.g., behavioral consultation; conjoint behavioral consultation) to bring about change directly in the consultees and indirectly in clients. The second tier concerns the treatment plan developed by the consultation team (consultant, consultees) and put into place by the consultee (teacher and/or parent). To conclude that consultation was efficacious in bringing about behavioral change in a client (child), evidence of implementation fidelity at both tiers is necessary. Understanding the factors that contribute to implementa-

tion fidelity across both tiers is important as we move forward in designing, developing, and implementing consultation-based educational interventions. Further investigation of factors that contribute to fidelity of both tiers of the independent variable will advance our ability to determine what constitutes important consultation intervention elements and how specific components contribute to salient outcomes. As a first step, it is essential that reliable and valid methods are available to measure fidelity characteristics.

Measurement of Fidelity in Consultation Research

Methods for collecting data to accurately assess implementation fidelity are emerging (Klute, Moreno, Sciarrino, & Anderson, 2008; Sanetti & Kratochwill, 2009; Trivette & Swanson, 2008; Zvoch et al., 2007). However, there are myriad issues regarding fidelity measurement within consultation, resulting in a dearth of research-based measurement tools (Noell, 2008; O'Donnell, 2008). Lack of clarity is evident around several issues, including (a) the identification and determination of plan components that predict treatment outcomes, (b) the relative weighting of specific plan components, (c) the sensitivity of measures to assess meaningful treatment components (i.e., those that contribute uniquely to treatment effects), (d) the appropriate metric that relates empirically to outcomes (e.g., item level, global score, or consistency in implementation over time), (e) the feasibility of assessment procedures (number of measurements required and ease of completion), (f) the scope of measurement necessary to capture a representative sample of treatment implementation, (g) the match of measurement tool to intervention components, (h) the reliability across measurement sources, and (i) the validity considerations. Each of these represents meaningful considerations that, if attended to in a systematic way, could contribute significantly to conceptual and empirical advances in fidelity research.

The process for determining treatment fidelity generally comprises three steps. The first involves identifying possible indicators or critical components of an intervention (i.e., specifying fidelity criteria). This critical step, based on expert consensus or the existence of evidence-based strategies, is often overlooked in intervention fidelity research. The second step is collecting data to measure the indicators, ideally through a multimethod, multiinformant approach. The third is to examine the indicators in terms of their reliability and validity (McGrew, Bond, Dietzen, & Salyers, 1994; Moncher & Prinz, 1991; Teague, Bond, & Drake, 1998).

Determining Fidelity Criteria

Procedures for determining fidelity criteria (i.e., steps that define treatment fidelity) primarily involve drawing from a specific intervention with demonstrable efficacy, effectiveness, or acceptability; gathering opinions from experts; or conducting qualitative research to ascertain the experiences and opinions of users (Moncher & Prinz, 1991). The most appropriate and feasible way to establish fidelity criteria is to examine the components of an evidence-based program that has proven successful, and develop a treatment manual to specify procedures and guide implementation (Bond, Williams, Evans, Salyers, Kim, & Sharpe, 2000). Ideally, fidelity criteria should be determined based on evidence that they contribute to treatment efficacy (i.e., they have treatment utility) or contribute unique variance to treatment effects. In this sense, fidelity criteria should have treatment validity and support a fidelity–outcome link (i.e., they should demonstrate that they contribute meaningfully to a treatment that is efficacious for a client). In some cases in consultation research, treatment steps are selected and defined given their objectivity (e.g., place name on paper; place materials in folder), without information on the predictive validity of each step to treatment outcomes. In these cases, high levels of fidelity can be observed; however, they provide little additional information regarding a treatment's efficacy. To date, little empirical research has been conducted on the effects of

fidelity criteria specified for measurement within consultation.

Measuring Fidelity

Quantification and measurement of intervention fidelity is possible via several means. The most common forms of assessment are ratings by experts or by participants in the intervention (Colton & Sheridan, 1998; Friesen et al., 2002). Ratings of implementation by experts or independent judges can be accomplished by relying on documentation or record review (e.g., permanent products yielding evidence of implementation; e.g., Mortenson & Witt, 1998; Noell, 2008), or through direct observations of implementation by treatment agents (live or via recordings; e.g., Jones, Wickstrom, & Friman, 1997; Mills & Ragan, 2000). Ratings by participants in the intervention are often in the form of self-report surveys or checklists by individuals delivering or receiving the intervention (e.g., Weiner, Sheridan, & Jenson, 1998).

In consultation, it is common for researchers to use self-report, permanent products, or direct observations to assess fidelity. *Self-report* measures are used to assess adherence of implementation as perceived by the parent or teacher. Typically, the steps comprising an intervention (i.e., fidelity criteria) are carefully delineated, and consultees record completion of the steps on an intervention-specific checklist. Thus, self-report assessments typically yield an estimate of consultees' adherence to or compliance with intervention implementation, computed as a percentage of steps completed. Benefits of self-report include its lessened reliance on extra human and material resources to collect information, ability to assess implementation using simple procedures in real time, and potential for inherently providing performance feedback to teachers (Sanetti & Kratochwill, 2008). As drawbacks, some researchers have suggested that self-reports result in overestimation of implementation by teachers (Lane, 2007; Wickstrom, Jones, LaFleur, & Witt, 1998); however, other researchers have found teachers to be accurate recorders. For example, Sanetti and Kratochwill (2009) found that agreement between self-report and permanent products is high.

Permanent products are used to assess intervention implementation via tangible evidence generated on intervention records or protocols. Examples of permanent product methods used to assess implementation fidelity are home–school notes (with intervention components clearly identified), charts, or tokens. They are completed daily by parents, teachers, and/or students and coded by observers on permanent product record forms. They provide an estimate of adherence, computed as percentage of intervention steps completed. There are many benefits of permanent products as a means of assessing treatment fidelity. First, permanent products offer a relatively simple measurement procedure that may not result in additional work or responsibility for the teacher or parent responsible for their completion. In some cases, interventions are structured in such a way that implementation results in a permanent record (e.g., charts or home notes), and may naturally provide important intervention information. Generally speaking, permanent products allow researchers to sample multiple occasions of an intervention with minimal reactivity by consumers (Sanetti & Kratochwill, 2008). However, interventions or intervention components do not always naturally result in a permanent product. Depending on the nature of the intervention, certain steps may be impossible to capture via a permanent product, such as those requiring teacher or parent judgment about behavioral compliance, quality of work, or appropriateness of social responses.

Direct observation as a method to assess treatment fidelity involves a trained and reliable individual, ideally independent from the consultation and intervention team, assessing direct, objective implementation of treatment plan components in naturalistic settings. This requires clear specification of both treatment plan components and observation sessions to capture direct evidence of implementation. Direct observations typically can be modified to address any intervention. Despite their apparent objectivity, direct observation is less common in intervention studies relative to other fidelity assessment methods. They can be very resource intensive, especially when additional

individuals (e.g., independent observers) are necessary to carry them out, or when numerous observations are required to capture implementation of various intervention steps. Observations may also produce reactivity among teachers and parents implementing the intervention; given the scope of certain interventions, they may not be representative of complex or broad-based intervention plans. Thus, direct observations may not produce valid fidelity results for all interventions equally.

Appropriate and sensitive measurement of implementation fidelity is among the greatest methodological challenges in intervention research (Noell, 2008). Each method has unique benefits; however, limitations inherent in each render them insufficient to assess intervention fidelity on their own. First, not all fidelity criteria are measurable with the same level of reliability, feasibility, or cost (Moncher & Prinz, 1991). Some criteria require judgments that are not objectively codified. Other intervention components are either lengthy or complex, thereby requiring extensive time, effort, and cost to capture all of its elements. As a result, in some cases all elements of an intervention program or model are not fully captured by the fidelity measurement, suggesting the need for multimethod approaches to fidelity assessment. However, interpretation of data collected from multiple sources is complicated, and to date there is no empirically derived approach to integrating and understanding data from multiple sources.

Assessing Reliability and Validity of Fidelity Measures

Although not common within the consultation literature, there are multiple methods for determining the psychometric qualities of fidelity measures. In a comprehensive review of the literature, Mowbray et al. (2003) identified several approaches to determining the reliability and validity of fidelity measures. The first is to examine reliability of measures across respondents by calculating inter-rater agreement through such means as percentage of agreement, intraclass correlation, or Pearson correlation coefficients. Second, the internal structure of the data can be determined empirically. Among the methods available for this type of analysis are internal consistency indices (e.g., Cronbach's alpha) or cluster factor analysis. Third, convergent validity, or procedures to determine agreement between two different sources of information about an intervention, has been used in studies on program implementation. For example, correlations between scores across fidelity rating scales (Lucca, 2000), and between permanent records and direct observations (Blakely et al., 1987), have been reported to provide evidence of convergent validity of fidelity measures. Finally, examining the relationship between fidelity measures and expected outcomes for participants is another approach to suggest validity of fidelity measurement. Ideally, these procedures will be used in studies wherein fidelity of implementation is of interest. However, the structure of fidelity measures used in practice settings does not always allow for systematic, rigorous psychometric tests, contributing to the importance of alternative means to determine their reliability and validity.

Exploratory Study: Psychometric Qualities of Fidelity Measures in Consultation Research

In the sections that follow we present an exploratory study that describes the development and field testing of intervention implementation fidelity measures used in consultation research, and we report preliminary psychometric properties of the fidelity measures across the two tiers of consultation interventions. This includes investigating the psychometric properties of measures used to assess intervention implementation by consultees (i.e., self-report, permanent products, direct observation) and consultation procedures used by consultants (i.e., direct observation). As this study is among the first to investigate psychometric qualities of fidelity measures used in consultation and was conducted in the context of a large-scale efficacy trial, challenges and issues pertaining to traditional methods of assessing reliability will be presented. Two exploratory questions concerning

preliminary psychometric properties of fidelity measures used in consultation are posed:

1. How reliable are each of the intervention fidelity measures (i.e., self-report, permanent product, direct observation) as assessed by intra- and inter-rater agreement and stability across time?
2. What is the reliability of independent observations of CBC fidelity (i.e., adherence to fidelity criteria) by consultants?

Methods

This study is part of a large randomized trial assessing the efficacy of CBC for addressing disruptive behaviors and the effects of CBC on parent–teacher relationships. It includes participants in the "CBC in the Early Grades Project," a federally funded intervention trial awarded to the first author.

Setting

The study took place in elementary schools in a large public school district and parochial schools in a moderately sized Midwestern city. Interventions were delivered in the classrooms and homes of students in an active treatment (CBC) condition.

Participants and Recruitment

Participants were 41 teachers, 101 parents, and 101 children in kindergarten through Grade 3.[1] Each classroom had at least 2 and up to 3 students participate in the study. Students' ages ranged from 5 to 9 years, with a mean age of 7 years of age; 78% of the students were male and 22% female. Sixty-five percent of students were reported to be White, non-Hispanic. Seventy-eight percent of parents characterized themselves as White. Fifty percent of parents reported acquiring less than a college degree. Sixty-four percent of families self-reported as meeting criteria to be considered living in poverty or low income (i.e., 25% and 39%, respectively). Forty-eight percent of the children received free or reduced-cost lunch at school.

Students were screened and selected for the project using the Systematic Screening for Behavior Disorders multiple-gate screening procedure (Walker & Severson, 1990). Students also qualified for the study if a teacher rated their behavior as severe and in grave need for additional intervention on a researcher-developed behavior severity scale (Glover, Sheridan, Garbacz, & Witte, 2005). Thus, a child could participate in the project if he or she was identified as a child who exhibited behavior concerns by the Systematic Screening for Behavior Disorders or the behavior severity scale. Once the students qualified, 2–3 students per classroom along with their guardians were selected at random to participate, and each classroom was then randomly assigned to an experimental or control condition. Only teachers, parents, and students in the CBC experimental condition are included in the present fidelity study.

Consultants in the study were seven clinicians trained in school or counseling psychology. All were female and self-reported as White, non-Hispanic. Consultants' average age was 25.43 ($SD = 2.23$); average years of graduate education completed was 2.57 ($SD = 1.81$). All but two (71%) had a master's degree at the start of the project; these two received their master's degree during the course of the study. Eighty-six percent were matriculated in a graduate training program at the time of the study; 71% were working toward a doctorate. Forty-three percent had been trained previously in CBC and underwent booster training for this study involving reading new CBC publications and engaging in group discussions regarding CBC service delivery in small group settings. Fifty-seven percent of the consultants had previous experience implementing behavioral interventions.

Procedures

CBC. CBC training was provided prior to initiation of CBC casework. This consisted of a 64-hr, criterion-based training program conducted over 4 weeks. Training was provided by project leaders, and included didactic instruction in CBC, readings on the theory and practice of CBC, readings about and tool kits for evidence-based behavioral interventions for disruptive behaviors, video demonstra-

tions, role-plays, self-monitoring, and direct individualized supervision. Furthermore, supervision was provided through CBC casework to maximize quality of service-delivery.

The structure for CBC casework was based on the approach described in Sheridan and Kratochwill (2008), with modifications to meet the demands of a large-scale randomized trial. Specifically, CBC implementation occurred in a series of stages, within a small group setting. Thus, within each classroom, a consultant met with the teacher and two to three parents for approximately four to five conjoint consultation sessions over approximately 8 weeks. All meetings occurred in teachers' classrooms and were approximately 45–60 min in length. The *Needs Identification/Needs Analysis (Building on Strengths*) phase involved collecting background information from various sources; discussing objectives; reviewing student, family, and school strengths; prioritizing one to two target behaviors per student; identifying and defining needs, settings, and goals; conducting functional behavior assessments; and discussing baseline data collection. Given the focus of the larger study, all behaviors identified for change were classified as disruptive behaviors. Beyond that, specific targets were selected and operationally defined for each student. Forty-two percent of targets were off-task behaviors. Following directions and classroom interference were each selected for 20% of the students, respectively. Fourteen percent of students had target behaviors that involved noncompliance, and 4% were aggressive.

The *Plan Development and Implementation* phase involved developing a plan to address student needs; training parents and teachers to implement the behavior plans; implementing the behavior plan; providing consultant support to maximize integrity via modeling, coaching, and performance feedback in the home and school setting; continuing to gather information about the child's behavior; and assessing intervention implementation integrity. Strategies used by consultants to maximize integrity included the development of manuals and scripted intervention plans, classroom and home-based observations with feedback, modeling, and recurrent phone/e-mail contacts for trouble shooting and support (Swanger-Gagne, Garbacz, & Sheridan, 2009). The *Plan Evaluation (Checking and Reconnecting)* phase involved discussing objectives; discussing progress made toward goals; evaluating the plans; and determining needs for plan continuation, modification, generalization, and/or fading.

Behavioral interventions and implementation planning. Behavioral interventions with clear evidence for effectively treating externalizing, disruptive behavioral concerns in children were developed to increase the content validity of the interventions and their subsequent implementation. A *CBC Behavioral Strategies Toolkit,* including teacher- and parent-friendly handouts based on materials and readings from *The Tough Kid Book* (Rhode, Jenson, & Reavis, 1992), *The Tough Kid Toolbox* (Jenson, Rhode, & Reavis, 1994), and *The Tough Kids Parent Book* (Jenson, Rhode, & Reavis, 2003), was developed for CBC consultants to use in their casework. The author of these materials (W. Jenson) provided consultation to the study's investigators to impart expert validity for the strategies chosen and the manuals developed supporting implementation of the evidence-based interventions. The CBC Behavioral Strategies Toolkit is organized by behavioral functions, including positive strategies and reductive techniques, and was used as the basis for individual behavior plan development. The Toolkit strategies were selected based on their empirical support (Stage & Quiroz, 1997). Strategies used in CBC cases to address disruptive behaviors were those that do the following: (a) *promote positive behaviors* through providing attention, rewards, and praise for compliance (Moore, Waguespack, Wickstrom, Witt, & Gaydon, 1994; Van Houten & Nau, 1980; Wolfe, Boyd, & Wolfe, 1983); (b) *reduce inappropriate behaviors* by setting limits and establishing consequences (McMahon & Forehand, 2003; White & Bailey, 1990; Witt & Elliott, 1982); (c) *provide proactive support* to guide students' behaviors (precision commands; skill training); and (d) *promote home-*

school communication through home notes or other consistent means (McCain & Kelley, 1994; Taylor, Cornwell, & Riley, 1984).

Efforts were made to standardize behavioral interventions across participants as much as possible. Specifically, standard behavior plans and manuals were developed for several evidence-based strategies contained in the CBC Behavioral Strategies Toolkit (e.g., chart moves, mystery motivators, grab bags). Individuation occurred at the level of specific reinforcers, schedules of reinforcement, and other idiosyncratic elements of individualized plans. Plans were clearly specified in terms of steps (criteria) that defined accurate implementation. Criterion-based fidelity checklists (i.e., self-report forms) were developed and shared with parents and teachers, and the self-report forms served as the basis for permanent product and direct observation report forms. Standardization across students was ensured further by including, for each individualized behavior plan, a motivation component (e.g., token economy, reward menu) and a home–school communication system to support continuity, cooperation, and support between parents and teachers outside of CBC meetings. The consultant scheduled additional contacts with parent(s) and the teacher outside of CBC sessions as needed to further specify details of the behavior plan (e.g., specific items for the reward menu). Consultants also provided additional training related to behavior plan implementation in the home and classroom settings during the treatment plan implementation stage.

Scale Development and Fidelity Assessments

The use of multimethod, multisource, multisetting measures to assess fidelity within consultation and intervention research is rare (Perepletchikova, Treat, & Kazdin, 2007). In the present study, we used three methods to assess fidelity: permanent products, self-report, and direct observation. Each measure yields an estimate of adherence of intervention implementation, computed as the percentage of fidelity criteria met by parents and teachers.

Fidelity criteria were derived by consultants, parents and teachers collaboratively during CBC meetings. Each intervention contained 3–12 criteria (steps) that defined accurate implementation (e.g., provided sticker for meeting playground goal; reviewed and signed home note), and each step was scored as "Yes" (step completed), "No" (step not completed), or "NA" (not applicable, in cases where child was not present, child failed to adhere to intervention demands, or other circumstances beyond the consultee's control). The criteria (steps) of each intervention were used in the development of the intervention implementation integrity measures. Specifically, they were translated onto three forms that had an identical structure but unique purposes: a *Parent and Teacher Self-Report Plan Summary forms*, *Home and School Permanent Product Report forms*, and *Classroom Observation Report forms*.

Self-Report Plan Summary Form. Self-report measures were used to assess parents' and teachers' self-recorded adherence of implementation, and yielded estimates of fidelity computed as percentage of steps of an intervention completed by consultees. The Self-Report Plan Summary forms contained a list of fidelity criteria that defined implementation steps. They were developed by the consultant during the plan development meetings or immediately thereafter, and reflected all of the steps agreed on by parents and teachers (i.e., they captured all intervention components). They were completed by parents and teachers daily while the intervention was in place. Teachers and parents received the Self-Report Plan Summary forms at the intervention planning meetings and were asked to self-record completion of all steps of plan implementation on a daily basis. On average, parents completed self-reports over 12.63 days (median = 11; range = 1–34; SD = 7.94); teachers completed an average of self-reports across 10.97 days (median = 11; range = 1–20; SD = 5.03).

Permanent Product Record Form. Interventions were structured in a way that routinely yielded a permanent record of implementation through the use of charts, record

forms, self-monitoring sheets, home notes, or other tangible records. Parents and teachers used the permanent products daily as a part of the intervention process. Fidelity criteria defined on the self-report forms were used by research assistants to develop Permanent Product Record forms. Specifically, items from the self-report forms that were observable on permanent products were listed on the Permanent Product Record form (e.g., received sticker for work completed, signed home note daily). Relative to the number of actual intervention steps within a case as defined on self-report forms, the average percentage of steps observable was 44% (range = 22%–88%; *SD* = 18%) for home permanent products, and 52% (range = 20%–100%; *SD* = 21%) for school permanent products.

Consultants collected intervention permanent products (e.g., home notes, charts) from consultees on a weekly basis. Research assistants documented the fidelity with which the consultees implemented the intervention by recording steps noted on the permanent products onto the record form. Permanent products were structured differently across cases (i.e., some were completed on a daily basis, such as home notes; others covered multiple days, such as weekly goal charts). Thus, there was variability in the number of permanent products possible across cases. An average of 12.01 school permanent products were collected across cases (median = 13; range = 1–20; *SD* = 4.89); similarly, an average of 13.26 home permanent products were collected (median = 14; range = 1–26; *SD* = 5.87).

Direct Observation Record Form. Direct observations of teachers implementing interventions were conducted by consultants over 4 weeks of intervention implementation. Fidelity criteria as defined on the self-report forms were used by research assistants to develop Direct Observation Record forms. These forms contained items on the self-report forms that were observable by an individual other than the teacher, including behaviors of treatment agents or environmental conditions (e.g., provided praise following desired behavior, established quiet work space for student). Relative to the number of actual intervention steps within a case as defined on self-report forms, the mean percentage of steps observable via direct observation was 74% (range = 33%–100%; *SD* = 19%).

Observation times were selected collaboratively by teachers and consultants, to be commensurate with periods of the day when interventions were being implemented. Between one and five classroom observations were conducted per case (*M* = 2.89; median = 3; *SD* = 1.17). Teachers were not aware that consultants were conducting formal observations of treatment fidelity. Because it was common for consultants to spend time in classrooms, their presence likely did not cause a reactive effect. No structured observations were conducted in homes.

CBC Objectives Checklists. Adherence with which consultants followed the objectives of CBC was assessed with *the CBC Objectives Checklists. The* checklists have been used in previous research (Sheridan et al., 2001) and were adapted to meet the unique circumstances presented by the large-scale randomized trial (i.e., small group consultation format). Each CBC interview consisted of specific objectives defining accuracy of delivery by consultants. The Needs Identification/Needs Analysis (Building on Strengths), Plan Development and Implementation, and Plan Evaluation (Checking and Reconnecting) Interviews consisted of 20, 10, and 10 objectives, respectively. Trained, independent coders listened to the digitally recorded interviews and coded the presence of each objective on the checklist.

Data Analysis

Various analyses were performed to explore the psychometric qualities of the fidelity measures (Mowbray et al., 2003). Procedures used to address each research question follow.

Research Question 1: Reliability of Fidelity Measures. Various indices of reliability were used to determine consistency of the various fidelity measures. First, exact agreement between two raters coding adherence of fidelity criteria (steps) as indicated

on Permanent Product Record forms was computed. Because averages are affected by scores at the extremes, and do not adequately represent tendencies when the majority of data are skewed in a particular direction (in this case, negatively skewed with few scores different than 100%), the median is used as the representative metric of agreement. Second, inter-rater reliability for the permanent product measure of home and school intervention integrity was computed using intraclass correlation coefficients (ICCs) from a one-way random effects model, where cases were considered random effects. Intraclass correlation values are interpreted as the percentage of the variability in fidelity scores that are caused by the differences across the cases that were rated, and controls for chance agreement. ICC is used to estimate the relationship among variables within a common measurement class, as compared to procedures (e.g., Pearson product) that measure the relationship of variables representing different classes (e.g., height, weight; McGraw & Wong, 1996).

To determine the stability of each measure over time (i.e., permanent product ratings at home and school; self-report measures completed by parents and teachers; direct classroom observations), standard deviations of fidelity scores across assessment or time points within a case were computed. Because we were concerned with identifying the degree to which each measure was stable across measured time points within each case, we used standard deviations as an estimate of the dispersion (or spread) of the values for each fidelity measure (i.e., self-report, permanent product, direct observation). The lower the standard deviation, the less variability there was around a common overall fidelity score. Furthermore, a high percentage of cases with standard deviations equaling zero suggested that fidelity was being measured consistently over time.

Research Question 2: Reliability of CBC Adherence Ratings. To measure reliability of independent observations assessing CBC adherence to procedures, the percentage of exact agreement and intraclass correlations was computed. Procedures similar to those used to address the reliability of permanent product data (see data analysis for Research Question 1 earlier) were used.

Results

Descriptive statistics for implementation of intervention steps are reported in Table 1. In general, mean levels of fidelity were high across sources and settings (range = 81.05%–98.57%). The lowest percentage of fidelity of intervention implementation was rated by parents in their self-reports of implementation; the highest was indicated in the ratings of permanent products collected from teachers. Because of the skewness in the data, with many fidelity scores at the high (close to 100%) end, the median rather than the mean may be a more representative metric of fidelity. The median estimate of implementation fidelity across settings and methods is 100%.

Research Question 1: Reliability of Measures

Measuring the reliability of each measure occurred in a variety of ways (i.e., exact agreement, intraclass correlations, and standard deviations). All (100%) permanent products were coded for evidence of intervention implementation fidelity by two trained observers. The median exact agreement was 88.68% for home-based and 98.57% for school-based permanent products. Inter-rater agreement for permanent products was further estimated by computing ICCs. ICC values can be interpreted as the percentage of the variability in integrity scores caused by the differences across the cases that were rated; the higher the ICC, the more one can interpret variability from cases rather than raters. In this study, ICCs were higher for ratings of school- rather than home-based permanent products. Specifically, an ICC of 0.986 at school indicates that 98.6% of the variability in fidelity scores on permanent products at school is from the difference across the cases that were rated, suggesting a strong degree of reliability when using only one rater. Approximately 72% of the variability of home-based permanent prod-

Table 1
Multisetting, Multimethod Intervention Fidelity Data[a]

Statistic	Self-Report	Direct Observation	Permanent Products
School intervention integrity[b]			
N[c]	78	89	70
Median	100%	100%	100%
Mean	92.15%	87.06%	98.57%
SD	16.82	20.16	11.95
Minimum	16.67	0	0
Maximum	100%	100%	100%
Skewness	−2.84	−1.85	−8.06
Kurtosis	8.76	3.79	66.12
Home intervention integrity[b]			
N[c]	49	NA[d]	53
Median	100%	NA	100%
Mean	81.05%	NA	87.82%
SD	27.42	NA	29.52
Minimum	0	NA	0
Maximum	100%	NA	100%
Skewness	−1.62	NA	−2.38
Kurtosis	2.26	NA	4.33

[a] The integrity value at each time point for each case was calculated. For each case, we then calculated a median value across the integrity values. These values were then summarized across cases to compute median, mean, and *SD*.
[b] School intervention integrity data were collected from teachers in the case of self-report and permanent products, and from consultants in the case of direct observations. Home intervention integrity data were collected from parents.
[c] *N* = sample size for which respective fidelity measurement is available. The availability of respective fidelity measures varies across cases.
[d] NA = not applicable.

uct ratings is from differences in variability across cases (ICC = 0.719). As expected, reliability increases with two raters, with ICCs estimated at 0.993 for school-based permanent products and 0.836 for home-based permanent products for two raters. Thus, 99.3% and 83.6% of the variability in fidelity scores was from variance across cases when coded by two individuals for school and home permanent products, respectively.

Standard deviations provide a metric of the stability of measurement present in self-reports, permanent product ratings, and direct observations collected over time. Results are in Table 2. The median standard deviation across the time points ranged from 0 (for school-based permanent products) to 14.86 percentage points (for parent self-reports). This suggests that across the measurement occasions, one can loosely expect fidelity scores to vary approximately ± 15 percentage points. It is noteworthy that across measures, between 28.3% and 63.2% of cases had standard deviations of 0, indicating that these cases received the exact same and consistently high fidelity value (percentage of fidelity criteria met) across all of the measured time points.

Research Question 2: Reliability of CBC Adherence Ratings

The degree to which consultants adhered to the objectives of CBC on the CBC Objectives Checklist was assessed. Thirty-three percent of CBC interviews were coded for fidelity purposes; of these, 33% were

Table 2
Descriptive Statistics of Standard Deviations (*SD*)[a] Across Time Points

Statistic	Parent Self-Report	Permanent Product—Home	Teacher Self-Report	Consultant Direct Observation	Permanent Product—School
N[b]	46	50	76	75	68
Mean	17.62	19.00	13.14	12.28	10.34
Median	14.86	12.14	10.84	9.62	0.00
Minimum	0.00	0.00	0.00	0.00	0
Maximum	57.74	52.22	40.82	57.74	54.77
Percentage of cases where *SD* for measure = 0	28.3	42.0	34.2	44.0	63.2

[a] Standard deviation for each case across time points (for cases with 2 or more time points) were calculated; values represent a summary of standard deviations across cases.
[b] Values are calculated on cases with two or more observed time points; thus, *N* values (representing number of cases) vary slightly from those found in Table 1.

coded by two trained raters to assess CBC process integrity. Across coders, an average of 96% of CBC objectives was identified as delivered by consultants (range = 80%–100%; *SD* = 0.08). Raters had exact agreement with the total number of objectives met across 84% of the interviews coded in common. The intraclass correlation was .515 for one rater; it increased to .680 for two raters. That is, approximately 52% of the variability of CBC fidelity was from differences in variability across cases; 68% of the variability in fidelity scores was from the variability across cases when coded by two individuals. The relatively low ICC is likely caused by the extremely narrow range of scores across coders and the limited variability in their ratings relative to one another.

Discussion

This study was the first to investigate the psychometric qualities of fidelity measures used in consultation research. It provides a summary of the fidelity of implementation across the two tiers of consultation: behavioral interventions implemented by parents and teachers, and delivery of CBC services by consultants. Traditional psychometric methods were used to assess the reliability of intervention permanent products and CBC objectives met by consultants. Standard deviations were used as an exploratory method to report on the stability of fidelity measures over time. In general, high levels of intervention integrity were noted across multiple methods and sources. In addition, a consistently high degree of intervention integrity was evident regardless of the method used and source. Although limitations in our data precluded us from conducting traditional tests of validity, the observed congruency of results across methods of assessment may serve as a proxy for convergent validity. Items included as fidelity criteria in our assessments were derived from the evidence-based literature and expert consultation, chosen based on their presumed predictive validity and not simply their ease of observation. This approach to fidelity assessment added to the validity of our measures.

This study yields encouraging preliminary findings for the psychometric adequacy of fidelity measures used in consultation, and in particular, permanent products. Because of inherent challenges associated with assessing psychometric qualities of common forms of fidelity assessment in consultation (i.e., self-report, direct observation), most information is available for the reliability of permanent products. Specifically, it is not possible to assess inter-rater reliability of self-report measures,

and the costs associated with having more than one observer in classrooms to determine interrater reliability of observations is prohibitive in most practical situations. Information on permanent products, however, is encouraging. ICCs suggest that permanent products can be coded reliably to determine some aspects of intervention integrity. However, it is important to recognize that permanent product data are limited in what they are able to capture. In our study, only between 44% and 52% of items (fidelity criteria) were able to be captured on permanent products at home and school, respectively. Only concrete steps or components of interventions are observable on permanent products and coded by independent raters. The same can be said for direct observations of intervention implementation in classroom settings; direct, independent assessments are possible for limited aspects of intervention implementation, covering a narrow time frame and scope of setting. In the current study, only 74% of items were observable through direct observations in classrooms. Thus, the findings herein represent certain features of the interventions only, and the representativeness of the results as recorded via permanent products and direct observations may be considered suspect. Continued efforts may uncover ways to operationalize components of behavioral interventions that are challenging to observe in live situations and contexts.

Our observations of overall fidelity ratings, based on the various measures collected within each case, suggest a great deal of similarity in estimates yielded by self-reports, permanent products, and direct observations (see Table 1). Estimates based on median percentages, which provide the most accurate measure of central tendency given the skewness in the data, yield identical estimates of fidelity (100%) across methods and sources. Mean ratings of fidelity differed somewhat, but not substantively (range = 81.05% on parent self-reports to 98.57% on school permanent products).

Whereas it is the case that the strength of measures used to assess fidelity and the resulting global scores are determined in part by the consistency of its items (i.e., fidelity criteria or intervention steps), we expect that the relationship between implementation fidelity and intervention outcomes is more a function of adherence to meaningful intervention protocols *over time* (i.e., dosage). Furthermore, the sensitivity of items (i.e., fidelity criteria) and their utility in predicting outcomes (i.e., validity) has not been determined in previous fidelity research. Many studies of fidelity that use a common, standard measure contain items because of their objective rather than functional or predictive nature (e.g., place name on paper, place papers in folder). Our measure tapped only items we believed to be operative in altering student behavior response, and thus varied somewhat across cases. This is certainly an area in need of further research.

Limitations of Study

Despite the contributions of this study, its exploratory and nonexperimental framework poses important limitations to our ability to draw conclusive inferences about the psychometric qualities of fidelity measures as used in consultation casework. These include issues with the sample from which data are available and general psychometric issues.

Sample. The sample for this study was drawn from a larger randomized trial assessing the efficacy of CBC for addressing disruptive behaviors among students in kindergarten through Grade 3. The number of participants involved in this study of fidelity was not equivalent to the total number available in the larger study. Specifically, whereas we were able to collect classroom-based fidelity data in a generally routine fashion, home-based fidelity data were available for only approximately 50% of all cases. This may have created a selection bias in that the parents returning the self-report and permanent product measures may have been those implementing interventions with greatest fidelity. It is not possible to corroborate a speculation regarding the behaviors of parents for whom no data are available (i.e., we cannot infer that integrity did or did not occur for these parents). Because no attempt to link between fidelity and child outcomes was attempted, the practical importance

of this issue is beyond the scope of this study. However, this sampling issue may have created a problem with restriction of range in our data, limiting our ability to perform certain analyses.

In addition to a select sample, incomplete or missing data within and across participants introduced variability in the amount of data generated and available across cases. Specifically, there were differences in the manner in which interventions were established, despite efforts to standardize them based on behavioral function. For example, some cases yielded permanent products that were completed and scored on a daily basis; others required recordings across days for completion. Likewise, differences in return rates across participants were evident, resulting in different amounts of data constituting overall levels of fidelity. For example, in some cases very few self-reports were available, whereas with others, self-report data were available covering several weeks. Similar issues are evident with permanent products, and to a lesser extent, direct observations. It is not possible to determine the absolute number of fidelity assessments expected per participant because the length of interventions and fidelity assessments was not standardized. That is, the researchers did not specify how many permanent products or other measures were expected per case, resulting in a range in how many and how consistently measures were provided. In addition, to compute standard deviations, cases for which fewer than two measures were available were excluded. It is possible that these cases were ones for which fidelity was lower than cases where several measures were available. Whereas these issues may be considered instances of missing data, the descriptive nature of our analyses do not allow for formal statistical correction (e.g., imputation, full information maximum likelihood; Enders, 2001). Future research using rigorous experimental designs allowing for data analytic approaches to formally handle missing data will be necessary in future research.

Psychometric issues. The select sample from which these data were drawn resulted in high ceiling effects for all fidelity measures. Integrity levels in this investigation were not manipulated experimentally, and the present study was not designed to achieve variability in the measurement of fidelity. Rather, data were part of a larger clinical trial and we used naturalistic adherence to fidelity criteria as the variable of interest. In fact, efforts were made to achieve high levels of fidelity and observe minimal variability in the implementation of behavioral interventions and consultation practice; therefore, the restriction of range and negative skewness in our data represented a positive circumstance in the larger test of intervention efficacy. However, the lack of variability in fidelity scores across time points on all measures resulted in a restriction of range. Specifically, there were many instances with scores at the high end of the distribution (nearing 100% fidelity); ceiling effects created a situation wherein computation of an alpha coefficient (which relies on a range of or variability within the data) was not possible. Future research is needed with the express purpose of developing reliable and valid tools for use in consultation research, thereby allowing for more rigorous examination of fidelity at many levels.

Psychometric limitations. Several issues inherent in our data set precluded our ability to use conventional approaches to investigate the psychometric qualities of fidelity measures. First, consistent with the idiosyncratic nature of behavioral consultation and CBC, interventions were individualized to address specific, within-participant issues (Mowbray et al., 2003). Although similar features were common across intervention plans (i.e., motivation, communication, and functional components were incorporated into all interventions), the individualization that occurred as a function of consultation resulted in different fidelity criteria (items) across cases. This, in turn, disallowed us to compute a Cronbach's alpha for an overall index of internal consistency. Second, cases presented a range of time points on which the fidelity measures were collected. Thus, listwise deletion procedures deleted all cases in which the

data were incomplete or where fewer than the maximum number of measurements was collected. This resulted in a sample size that was insufficient to compute Cronbach's alpha for the fidelity measure.

Standard deviation of scores across time points was used as a nontraditional means of reporting stability of our fidelity measures. In this regard, we were concerned with the percentage of cases for which the standard deviation equaled zero as a reflection of the consistency in ratings over time. Findings suggest that between 28% and 63% of the measures demonstrated complete stability over time (i.e., $SD = 0$); however, because this use of standard deviation is somewhat unique, there are currently no interpretive guidelines suggesting "acceptable" stability levels. Nevertheless, we recognize the inherent limitations with data that are skewed, and believe this approach allowed us to take advantage of standard deviation as a useful estimate of variability without violating assumptions of normality.

To determine convergent validity, correlations among three measures of intervention fidelity at school (self-report, permanent product, direct observation) and two measures of intervention integrity at home (self-report, permanent product) were attempted, but unattainable for several reasons. First, as already mentioned, high ceiling effects were evident for all measures, such that the vast majority of intervention fidelity measures yielded rates of 100%. Second, although close to perfect correspondence was attempted, different fidelity criteria were evident across measures and methods. For example, certain items recorded on self-report forms were not observable on permanent products or in direct observations, and thus were omitted from that measure. The lack of perfect congruence made efforts at aligning or investigating direct associations among measures difficult. Finally, few measures were completed at the same exact time points (e.g., direct observations, self-reports), making direct, point-to-point convergence impossible to assess. More precise validation of self-report, permanent product, and direct observation methods/forms for assessing intervention integrity will be necessary to advance the field in specific, necessary directions.

It is noteworthy that many of these limitations are recognized barriers to measuring implementation fidelity (e.g., Noell, 2008). Specifically, there was a lack of standardization for intervention plan components and difficulty in determining the relative "weight" (i.e., importance) of each component. In addition, different criteria were evident across measures and methods. Furthermore, the lack of congruence across cases with regard to fidelity criteria and number of measured time points represents a need to more clearly operationalize implementation fidelity, including standardizing measures, in consultation research (Noell, 2008).

Future Research Directions

The limitations inherent in this investigation shed light on essential areas in need of research attention. It is clear that fidelity research is "coming of age" and is no longer possible only within the context of larger intervention studies. Research that treats fidelity of implementation as an independent variable, with clear specification and experimental manipulation, will allow for the determination of many important directions in consultation research. As the need for this type of research increases, the importance of understanding methods to assess fidelity reliably and meaningfully will magnify. Clearly, the confidence in decisions related to consultation process and intervention outcomes is only as plausible as the clarity with which we can define and ensure implementation of active treatment components. Identification of "active" (i.e., functional and predictive) components of interventions that translate into fidelity criteria requires careful, sound measurement.

A common fidelity indicator assessed in consultation research is adherence (e.g., documentation of the delivery of consultation objectives or intervention steps as designed). However, assessment of adherence to criteria as a sole metric of fidelity represents a narrow conception of the construct. Fidelity of implementation in intervention studies is actually a

multidimensional construct and can be characterized along five dimensions: *adherence, dosage, quality of program/intervention delivery*, *participant responsiveness*, and *program differentiation* (Dusenbury et al., 2003; O'Donnell, 2008). Adherence is conceptualized as the implementation of intervention strategies as designed by program developers. Dosage is the overall amount of intervention delivered to participants (also considered "strength" of the intervention; Sechrest & Yeaton, 1981). The quality of intervention delivery is a step beyond adherence indicating the quality or effectiveness with which intervention strategies are delivered by practitioners (or the "competence" of treatment agents to deliver treatments; Cordray & Pion, 2006). Participant responsiveness indicates the participants' level of engagement in and receptiveness to intervention programming. Finally, program differentiation indicates whether the characteristics of the intervention distinguish treatment from control groups during the implementation of the intervention in studies evaluating the efficacy of interventions. It is uncommon in the consultation literature for issues regarding dosage (e.g., number of hours of consultation or overall amount of intervention components delivered), quality (i.e., appropriateness of parent/teacher implementation or effectiveness of consultant actions), participant uptake (i.e., amount of intervention received or responded to by the consultee or client), or differentiation among programs (e.g., consultation vs. other direct or indirect services) to be targeted, measured, or manipulated in consultation research. In this vein, consultation researchers lag behind those in the prevention sciences.

Research is needed to determine the cost–benefit ratio of various assessment procedures. The econometric aspects of fidelity assessment are real and represent practical issues in the inclusion of measures in consultation practice and research. The question of "how often" or "how much is enough" needs to be examined. Measures can be costly in terms of scale development, parent and teacher time, observer training, and personnel to carry out independent assessments. There is a need to determine critical points at which assessment fails to provide unique intervention information.

Fidelity as a construct is useful for a number of reasons. Among the most important is the information we can obtain to enhance our understanding of treatment effects. A fidelity measure has particular treatment validity if it can help explain the manner in which an intervention exerts its effects and rule out alternative explanations (threats) to the validity of conclusions regarding an intervention's efficacy. However, it is best conceptualized as part of a comprehensive approach to intervention research. Specifically, articulation of many additional variables must also be specified to maximize the utility of fidelity measures, including the theoretical formulation undergirding the treatment; documentation of competing or confounding variables and nonspecific effects (e.g., expectancy, placebo effects); description and measurement of the counterfactual and control conditions; and hypothesized strength of the treatment (Cook & Campbell, 1979; Cordray & Pion, 2006). All of these are worthy aspects of fidelity to be included in future consultation research.

Experimental attention in the area of fidelity of implementation will provide essential information on the utility of specific methods for promoting intervention implementation integrity. Furthermore, the relationship between fidelity and important research outcomes (i.e., at the child and parent/teacher, and system levels) can be discerned only with evidence of fidelity. It is likely that intervention implementation fidelity acts as a mediator of consultation effects on identified outcomes, and experimental research is necessary to explore this specifically. Careful, systematic scrutiny of the intervention and aspects of implementation will increase researchers' abilities to uncover unique, important components of interventions in ways that expand beyond simply assessing adherence and dosage. That is, elements of quality, and not simply quantity, can be discerned, with the effect of determining precisely what works, for whom, how, and under what conditions.

Footnotes

[1]The numbers reported here reflect the total number of participants from the larger study for whom some data are available. Different measures are available for different participants (e.g., for some, self-report data may be available; for others, permanent product data may be available).

References

Bergan, J. R., & Kratochwill, T. R. (1990). *Behavioral consultation in applied settings.* New York: Plenum.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., et al. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology, 15,* 253–268.

Bond, G. R., Williams, J., Evans, L., Salyers, M. P., Kim, H. W., & Sharpe, H. (2000). *PN-44-psychiatric rehabilitation fidelity toolkit.* Cambridge, MA: Human Services Research Institute.

Bronfenbrenner, U. (1979). *The ecology of human development: Experimental by nature and design.* Cambridge, MA: Harvard University Press.

Colton, D., & Sheridan, S. M. (1998). Conjoint behavioral consultation and social skills training: Enhancing the play behavior of boys with attention deficit-hyperactivity disorder. *Journal of Educational and Psychological Consultation, 9,* 3–28.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Chicago, IL: Rand McNally.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18,* 23–45.

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3,* 269–276.

Drake, R., Goldman, H., Leff, H., Lehman, A., Dixon, L., Mueser, K., et al. (2001). Implementing evidenced-based practices in routine mental health service settings. *Psychiatric Services, 52,* 179–182.

Durlak, J. A. (1998). Why program implementation is important. *Journal of Prevention & Intervention in the Community, 17,* 5–18.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18,* 237–256.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370.

Erchul, W. P., & Schulte, A. C. (1996). Behavioural consultation as a work in progress: A reply to Witt, Gresham, and Noell. *Journal of Educational and Psychological Consultation, 7,* 345–354.

Frese, F. J., Stanley, J., Kress, K., & Vogel-Scibilia, S. (2001). Integrating evidence-based practices and the recovery model. *Psychiatric Services, 52,* 1462–1468.

Friesen, B. J., Green, B. L., Kruzich, J. M., & Simpson, J., et al. (2002). *Guidance for program design: Addressing the mental health needs of young children and their families in early childhood education settings.* Available from Portland State University, Research & Training Center on Family Support and Children's Mental Health Web Site, http://www.rtc.pdx.edu/pgProjGuidance.php

Glover, T., Sheridan, S. M., Garbacz, S. A., & Witte, A. (2005). *Behavior severity, behavior frequency and need for intervention screening tool.* Unpublished scale.

Guli, L. A. (2005). Evidence-based parent consultation with school-related outcomes. *School Psychology Quarterly, 20,* 455–472.

Jenson, W. R., Rhode, G., & Reavis, H. K. (1994). *The tough kid tool box.* Longmont, CO: Sopris West.

Jenson, W. R., Rhode, G., & Reavis, H. K. (2003). *The tough kids parent book: Practical solutions to tough childhood problems.* Longmont, CO: Sopris West Publishing.

Jones, K. M., Wickstrom, K. F., & Friman, P. C. (1997). The effects of observational feedback on treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 12,* 316–326.

Klute, M. M., Moreno, A. J., Sciarrino, C. A., & Anderson, S. (2008). Fidelity and implementation of "Learning through Relating," a pre-literacy and social communication curriculum for infants and toddlers. In W. DeCourcey (Chair), *Head Start university partnerships: Curriculum development.* Paper presented as part of a symposium at the annual convention of the Head Start Ninth National Research Conference, Washington, DC.

Lane, K. L. (2007). Identifying and supporting students at risk for emotional and behavioral disorders within multi-level models: Data driven approaches to conducting secondary interventions with an academic emphasis. *Education and Treatment of Children, 30,* 135–164.

Lucca, A. M. (2000). A Clubhouse fidelity index: Preliminary reliability and validity results. *Mental Health Services Research, 2,* 89–94.

McCain, A. P., & Kelley, M. L. (1994). Improving classroom performance in underachieving preadolescents: The additive effects of response cost to a school-home note system. *Child & Family Behavior Therapy, 16,* 27–41.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

McGrew, J. H., Bond, G. R., Dietzen, L., & Salyers, M. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology, 62,* 670–678.

McMahon, R. J., & Forehand, R. L. (2003). *Helping the noncompliant child* (2nd ed.). New York: Guilford Press.

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning sys-

tem (ILS). *Educational Technology Research and Development, 48,* 21–41.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247–266.

Moore, L. A., Waguespack, A. M., Wickstrom, K. F., Witt, J. C., & Gaydon, G. R. (1994). Mystery motivator: An effective and time efficient intervention. *School Psychology Review, 23,* 106–117.

Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review, 27,* 613–627.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24,* 315–340.

Noell, G. H. (2008). Studying relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 323–342). Mahwah, NJ: Erlbaum.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33–84.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75,* 829–41.

Rhode, G., Jenson, W. R., & Reavis, H. K. (1992). *The tough kid book.* Longmont, CO: Sopris West.

Rossi, P. H., & Freeman, H. E. (1985). *Evaluation: A systematic approach.* Newbury Park, CA: Sage.

Sanetti, L. M. H., & Kratochwill, T. R. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4,* 95–113.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. *School Psychology Quarterly, 24,* 24–35.

Sechrest, L., & Yeaton, W. E. (1981). Assessing the effectiveness of social programs: Methodological and conceptual issues. In S. Ball (Ed.), *New directions for program evaluation: Assessing and interpreting outcomes* (pp. 41–56). San Francisco: Jossey-Bass.

Sheridan, S. M., Clarke, B. L., & Burt, J. D. (2008). Conjoint behavioral consultation: What do we know and what do we need to know? In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 171–202). Mahwah, NJ: Lawrence Erlbaum.

Sheridan, S. M., Eagle, J. W., Cowan, R. J., & Mickelson, W. (2001). The effects of conjoint behavioral consultation: Results of a four-year investigation. *Journal of School Psychology, 39,* 361–385.

Sheridan, S. M., & Kratochwill, T. R. (2008). *Conjoint behavioral consultation: Promoting family-school connections and interventions.* New York: Springer.

Sheridan, S. M., Kratochwill, T. R., & Bergan, J. R. (1996). *Conjoint behavioral consultation: A procedural manual.* New York: Plenum Press.

Stage, S. A., & Quiroz, D. R. (1997). A meta-analysis of interventions to decrease disruptive classroom behavior in public education settings. *School Psychology Review, 26,* 333–368.

Swanger-Gagne, M., Garbacz, S. A., & Sheridan, S. M. (2009). Intervention implementation integrity within conjoint behavioral consultation: Strategies for working with families. *School Mental Health, 1,* 131–142.

Taylor, V. L., Cornwell, D. D., & Riley, M. T. (1984). Home-based contingency management programs that teachers can use. *Psychology in the Schools, 21,* 368–374.

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity and assertive community treatment: Development and use of a measure. *American Journal of Orthopsychiatry, 68,* 216–232.

Trivette, C., & Swanson, J. (2008). Windows of opportunity: Treatment fidelity results from a capacity building model with Early Head Start home visitors. In W. DeCourcey (Chair), *Head Start university partnerships: Curriculum development.* Paper presented as part of a symposium at the annual convention of the Head Start Ninth National Research Conference, Washington, D.C.

Van Houten, R. V., & Nau, P. A. (1980). A comparison of the effects of fixed and variable ratio schedules of reinforcement on the behavior of deaf children. *Journal of Applied Behavioral Analysis, 13,* 13–21.

Walker, H. M., & Severson, H. H. (1990). *Systematic screening for behavior disorders.* Longmont, CO: Sopris West.

Weiner, R., Sheridan, S. M., & Jenson, W. R. (1998). Effects of conjoint behavioral consultation and a structured homework program on math completion and accuracy in junior high students. *School Psychology Quarterly, 13,* 281–309.

White, A. G., & Bailey, J. S. (1990). Reducing disruptive behaviors of elementary physical education students with Sit and Watch. *Journal of Applied Behavioral Analysis, 23,* 353–359.

Wickstrom, K. F., Jones, K. M., LaFleur, L. H., & Witt, J. C. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13,* 141–154.

Wilkinson, L. A. (2006). Assessing treatment integrity in behavioral consultation. *The International Journal of Behavioral Consultation and Therapy.* Retrieved December 14, 2008, from http://findarticles.com/p/articles/mi_6886/is_3_3/ai_n28460884/pg_9

Witt, J. C., & Elliott, S. N. (1982). The response cost lottery: A time efficient and effective classroom intervention. *Journal of School Psychology, 20,* 155–161.

Wolfe, V. V., Boyd, L. A., & Wolfe, D. A. (1983). Teaching cooperative play to behavior-problem preschool children. *Education and Treatment of Children, 6,* 1–9.

Zvoch, K., Letourneau, L. E., & Parker, R. P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation, 28,* 132–150.

Susan M. Sheridan, PhD, is Willa Cather Professor and Professor of Educational (School) Psychology at the University of Nebraska—Lincoln. She is the director of the Nebraska Center for Research on Children, Youth, Families and Schools and the National Center for Research on Rural Education. Her research interests include family–school partnerships, conjoint behavioral consultation, and behavioral/social skills interventions.

Michelle S. Swanger-Gagné, MA, is a postdoctoral fellow in Clinical Psychology at the University of Rochester Medical Center. She received her MA in educational psychology from the University of Nebraska—Lincoln. She is a doctoral candidate at the University of Nebraska—Lincoln and expects to receive her PhD in School Psychology from the University of Nebraska—Lincoln in the fall of 2009.

Greg W. Welch, PhD, is Research Assistant Professor in the Statistics and Research Methods unit in the Nebraska Center for Research on Children, Youth, Families and Schools. He received his PhD from the University of Pittsburgh and was a faculty member at the University of Kansas prior to coming to the University of Nebraska—Lincoln. His research interests include the development of structural equation modeling techniques and the use of these techniques in helping shape educational policy decisions.

Kyongboon Kwon, PhD, is a postdoctoral fellow at the Nebraska Center for Research on Children, Youth, Families and Schools at the University of Nebraska—Lincoln. She received her PhD in Educational Psychology in 2008 from the University of Georgia. Her research interests include children's peer group socialization and family–school interventions.

S. Andrew Garbacz, MA, is a doctoral candidate in the University of Nebraska—Lincoln School Psychology Program. His research interests include family–school partnerships within and across countries, and the delivery of functionally appropriate interventions in family-centered, consultation models.

Date Received: January 2, 2009
Date Accepted: August 8, 2009
Action Editors: Lisa Sanetti and Thomas Kratochwill ■