Section B

METHODOLOGICAL AND STATISTICAL FOUNDATIONS
Measurement plays an important role in intervention research. It also takes many forms. Sometimes it involves simple counting. For example, knowing the number of participants in a study is useful in evaluating the confidence that can be placed in the results. More frequently, measurement in treatment research is complex. For example, assessing the integrity of intervention implementation or quantifying how much participants improved as a result of the treatment both require the operationalization and scaling of complex, multidimensional constructs.

This chapter discusses a broad array of measurement issues in school consultation research, from the simple to the complex. To limit the chapter’s length and breadth, the choice of measurement topics included are those that are central to one of the most important challenges facing school consultation researchers—demonstrating that school consultation is an effective treatment that merits more widespread use and support. The increased focus on, and funding for, the establishment and use of “evidenced-based” or empirically supported treatments in schools (Feuer, Towne, & Shavelson, 2002; Gallimore & Santagata, 2006; Grissmer, Subotnik, & Orland, 2009) offers an unprecedented opportunity to further develop and promote consultation as a viable and cost-effective treatment. It has already yielded benefits in terms of rigorously conducted consultation studies (e.g., DuPaul et al., 2006; Fabiano et al., 2010; Jitendra et al., 2007; Sheridan et al., 2012). However, consultation researchers’ ability to further capitalize on this opportunity depends on researchers continuing
to meet the rigorous and evolving methodological standards that are part of the move toward evidence-based treatments in a number of human services fields.

The chapter begins with a brief summary of the evidence-based treatment movement and its roots in medicine. It then moves to a discussion of what should be measured in consultation research and how it should be done, focusing on three major aspects of school consultation research: (a) describing the participants, (b) documenting the consultation process, and (c) assessing client outcomes. The chapter concludes with recommendations related to increasing the rigor of measurement in school consultation research and reducing the extent to which measurement issues may distort or attenuate the findings supporting school consultation. In keeping with the emphasis on consultation as a treatment, the focus will be on research concerning triadic case-centered consultation in schools, where the consultant (typically a psychologist, counselor, or specialist) meets with a direct service provider (teacher) to address a concern about a client (child), with the primary focus on improving the functioning of the client rather than the consultee. Although our definition of triadic case-centered consultation includes consultation models that involve both a professional and parents as consultees (e.g., conjoint behavioral consultation; Sheridan & Kratochwill, 2008), we do not include parent consultation without the involvement of a school staff member in our definition, as consultation was originally conceptualized as a process involving two professionals (Caplan, 1970; Gallessich, 1982).

EVIDENCE-BASED INTERVENTIONS IN SCHOOLS

The research base that supports the use of psychological and educational treatments has come under increased scrutiny in recent years. As part of a broad-based effort to encourage the use of interventions that have been shown to be effective in methodologically sound studies, several groups have begun to systematically review the quality of scientific evidence for psychological and educational treatments, as well as the effectiveness of broader social policies (e.g., Center for Data-Driven Reform in Education, 2012; Chambless et al., 1998; Coalition for Evidence-Based Policy, 2012).

These groups’ activities are part of a larger movement that began in the 1990s in the United Kingdom with a focus on the evidence to support medical treatments (Spring, Pagoto, Altman, & Thorn, 2005). However, the roots of the movement can be traced to the 1970s, or even earlier (Claridge & Fabian, 2005). Initially termed “evidence-based medicine” (Chambless & Ollendick, 2001), the major goals of this movement are to make summaries of the evidence supporting various treatments easily available to clinicians and consumers, and to promote the integration of high quality research evidence and individual clinical expertise in clinical decision making (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996).

Within medicine, this movement has had a profound impact (Bernstein, 2004). Its most well-known outcome is the Cochrane Collaboration (www.cochrane.org/). This international organization, founded in 1993, uses meta-analysis to systematically review the effects of healthcare interventions. It then makes the results of these reviews available to professionals and consumers through the Cochrane Database of Systematic Reviews. To date, the Cochrane Collaboration has published more than 5,000 reviews of treatment research in specific areas.
Although the evidence-based movement began with a focus on medical treatments, as it gained momentum, other human service areas began to examine the research support for their practices (Spring et al., 2005). For example, in 1993, Division 12 (Clinical Psychology) of the American Psychological Association appointed the first of two task forces to review the evidence for a number of psychological interventions for adults and children (e.g., Chambless et al., 1998; Lonigan, Elbert, & Johnson, 1998; Spirito, 1999). Later, Division 16 (School Psychology) of the American Psychological Association joined with the Society for the Study of School Psychology to support the formation of the Task Force on Evidence-Based Interventions in School Psychology (Kratochwill & Stoiber, 2002; hereafter called the Task Force on EBI).

In 1999, the Campbell Collaboration, a sibling to the Cochrane Collaboration, was established “to help policymakers, practitioners, and the public make well informed decisions about policy interventions by preparing, maintaining, and disseminating systematic reviews of the effectiveness of social and behavioural interventions in education, crime and justice, and social welfare” (Davies & Boruch, 2001, p. 294). In 2002, the U.S. Department of Education partnered with the Campbell Collaboration and established the What Works Clearinghouse (WWC; http://ies.ed.gov/ncee/wwc/) with a mission to study and report on the effectiveness of educational interventions (WWC, 2011). As of mid-2012, the WWC had published reviews of the evidence to support over 500 educational interventions, as well as produced 16 practice guides summarizing the empirical support for practices in targeted areas of education (e.g., data-based decision making).

This chronology does not provide a complete list of organizations examining the research bases of medical, psychosocial, and educational practices, but illustrates the breadth of this movement. One result of the number of private and public agencies completing systematic reviews of research evidence has been increased scrutiny of the overall quality of treatment research methodology and reporting. For example, an outgrowth of the Cochrane Collaboration has been the Consolidated Standards of Reporting Trials or CONSORT statement (Schulz, Altman, Moher for the CONSORT Group, 2010). This document is a checklist of essential items that are to be included when reporting the results of randomized controlled trials. The document was developed by an international group of journal editors and investigators when it became apparent that incomplete reporting of the methodology used in published clinical trials was a major barrier to systematic reviews of the evidence for healthcare interventions (Moher et al., 2010).

Since their original publication, the CONSORT guidelines have now been revised twice, and their use appears to have improved the quality of reporting in published treatment outcome studies (Hopewell, Dutton, Yu, Chan, & Altman, 2010). Both the current CONSORT 2010 reporting template (Schulz et al., 2010) and the accompanying explanation and rationale for each of its components (Moher et al., 2010) are valuable resources for anyone designing research to investigate psychological or educational treatments.

Following in the footsteps of the CONSORT Group, the WWC recently issued the What Works Clearinghouse Reporting Guide for Study Authors (WWC, 2012) to help education researchers report their findings in a clear and complete manner. The more complete and uniform reporting encouraged in this document, in turn, facilitates
application of the agency’s methodological standards, as summarized in the What Works Clearinghouse Procedures and Standards Handbook (WWC, 2011). These procedures and standards are used to characterize the extent and quality of evidence supporting particular educational interventions.

Other groups have also articulated specific standards against which to judge the quality of the available research studies investigating different treatments. For example, the Task Force on EBI has drafted the Procedural and Coding Manual for Review of Evidence-Based Interventions (Task Force on EBI, 2008; hereafter the EBI Procedural and Coding Manual). This manual articulates specific guidelines for use in judging the strength of evidence to support psychological and educational treatments used in schools. Not surprisingly, measurement has been one focus of this increased scrutiny on treatment research methodology. Categorization, quantification, and description play a critical role in the design, implementation, and evaluation of treatment research. For example, a key domain in the EBI Procedural and Coding Manual (Task Force on EBI, 2008) is the quality of the measures used to assess outcomes. The careful scrutiny of measurement issues in treatment research appears merited. In an extensive methodological meta-analysis of over 16,000 psychological and educational treatment studies, Wilson and Lipsey (2001) found that how constructs were measured contributed as much to the variance in treatment effect sizes as what was measured.

In sum, there is an increasing emphasis on the use of evidence-based practices in psychology and education. This emphasis has resulted in stronger demand for high quality intervention research, and a key to high quality intervention research is adequate measurement of important study variables and components.

What Should Be Measured in School Consultation Research and How?

As noted above, when treatment literature has been systematically reviewed, it has become evident that not all the information needed to fully assess the quality of, and empirical support for, various medical and psychosocial treatments is available in published reports (Hagermoser Sanetti, Gritter, & Dobey, 2011; Moher et al, 2010; Perepletchikova, Treat, & Kazdin, 2007). Problems with the quality of reporting in the consultation research literature have also been noted (e.g., Reddy, Barboza-Whitehead, Files, & Rubel, 2000; Sheridan, Welch, & Orme, 1996). Given these findings, this section focuses on information that should be included in research reports and the measurement issues that arise in efforts to obtain this information.

In presenting what variables are important to include in school consultation research and how they should be measured, four sources that address issues related to conducting and reporting treatment research will be used as primary guides. These sources are: (a) the guidelines contained in the EBI Procedural and Coding Manual proposed by school psychology’s Task Force on EBI (2008); (b) Gresham and Noell’s (1993) chapter concerning issues in documenting treatment outcomes in consultation research; (c) Weisz, Doss, and Hawley’s (2005) methodological review and critique of the youth psychotherapy research; and (d) documents produced by the WWC to guide authors in reporting studies and reviewers in evaluating the
adequacy of studies, including the *What Works Clearinghouse Procedures and Standards Handbook* (WWC, 2011) and *What Works Clearinghouse Reporting Guide for Study Authors* (WWC, 2012). Commonalities among these sources are used as a basis for inferring current standards or best practices in treatment research, either because the sources have explicitly stated standards for reporting or evaluating treatment research (the EBI Procedural and Coding Manual and the WWW publications), or they have evaluated the quality of treatment research on a number of dimensions (Gresham & Noell, 1993; Weisz et al., 2005). These four sources were selected because of their influence on education policy and research (the WWC publications), their relevance to child treatment research (Weisz et al., 2005), or their direct relevance to school consultation research (Gresham & Noell, 1993; Task Force on EBI, 2008).

The discussion of what variables should be measured in school consultation research is divided into three subsections, each addressing a major area of treatment research design. First, key considerations in the description of participants are addressed. Second, measurement issues involved in documenting consultation process are discussed. Finally, issues relevant to the measurement of client outcomes are presented. For each of these topics, salient reporting or measurement issues from the four methodological sources (i.e., Gresham & Noell, 1993; Task Force on EBI, 2008; Weisz et al., 2005; WWC, 2011, 2012) are discussed. When appropriate, the state of practice in the school consultation research literature, as indicated by comprehensive reviews (e.g., Reddy et al., 2000; Sheridan et al., 1996), is compared to the state of practice in the youth psychotherapy research literature, as portrayed in Weisz et al.’s (2005) comprehensive review of this research area. Such a comparison is useful because it provides a normative perspective on the state of school consultation research.

Although these comprehensive reviews of the research methodology in consultation and psychotherapy are somewhat dated, more recent methodological reviews have focused on a single methodological topic in consultation or psychotherapy (e.g., Hagermoser Sanetti & Kratochwill, 2008; Perepletchikova et al., 2007), or reviewed studies from domains broader than triadic consultation or youth psychotherapy (e.g., Guli, 2005; Hagermoser Sanetti et al., 2011). However, these more recent methodological reviews suggest that many of the problems cited in earlier reviews have yet to be fully addressed in the consultation and treatment literature.

**Description of Study Participants**

A key component of any report of research is a description of the number and characteristics of the participants. Knowledge of sample size is important in assessing a study’s statistical power (Sink & Mvududu, 2010). A complete description of the participants is important in understanding the study’s generalizability or external validity (Task Force on EBI, 2008) and identifying potential variables that may moderate treatment outcomes (Weisz et al., 2005).

**Sample Size**

A count of participants would seem to be one of the simplest measurement tasks facing researchers. In fact, it seems so simple that it may be difficult to even think of sample size as a measurement issue in consultation research. However, a serious
problem found in much of the overall treatment literature is the failure to report changes in sample size from the beginning to the end of treatment, and to report the number of participants that have been included or excluded in calculations of treatment outcomes (Moher et al., 2010). When interventions take place across a significant time interval, it is likely that the number of participants remaining, as well as the number complying with the intervention protocol, will change over the course of the study. Under these circumstances, it can be difficult to determine how to count participants because the number may vary over time and based upon the criteria used to define study participation.

School consultation research presents particular problems in counting participants because continued study participation depends not only on the client’s continued willingness to participate in treatment, but also the consultee’s. For example, in the classic behavioral consultation study by Bergan and Tombari (1976), 806 cases were referred to consultants for psychological services. Of these, consultation was initiated in only 43% of the cases. Consultation proceeded to the plan implementation phase for 31% of the original cases. In only 30% of the original referral cases did problem solution occur. DuPaul et al. (2006), in their consultation study that spanned two school years, reported that for the 167 student participants, 48 of the students’ second-year teachers refused to participate in consultation. Drops in participants from referral to treatment completion, particularly if the treatment is extended, do not seem to be unusual in consultation research, but the implications of this phenomenon for assessing consultation’s overall efficacy or effectiveness are often not recognized.

Both the EBI Procedural and Coding Manual (Task Force on EBI, 2008) and the WWC Procedures and Standards Handbook (WWC, 2011) require an examination of study attrition by treatment group and the manner in which study attrition was handled in data analysis procedures. These standards suggest that it is important for consultation researchers to carefully track and report study sample sizes from initial entry into treatment through follow-up. Two useful resources for reporting sample size across the stages of a study are the Flow Diagram from the CONSORT 2010 guidelines (www.consort-statement.org/consort-statement/flow-diagram/) for randomized controlled trials (RCTs) and the Attrition Diagram for single-case studies included in the pilot standards for single-case designs that are part of the WWC Procedures and Standards Handbook (WWC, 2011).

When sample sizes differ across intervention phases, researchers must decide which participants to include in the calculation of study outcomes. In an “intention to treat” analysis, data from all of the participants originally assigned to a treatment condition are included in the analysis of treatment outcomes. In an “on-treatment” or “per protocol” analysis, only participants who followed the intended study protocol are included in the outcome analysis (Moher et al., 2010). If the total number of participants in a treatment study varies markedly by these two types of analyses, then the choice of analysis strategy also is likely to affect the odds of obtaining a statistically significant result and the magnitude of the effect reported.

Kratochwill, Elliott, and Busse’s (1995) report of a five-year evaluation of consultant training provides an example of how the choice of analysis and consequently who is counted as a participant in calculating outcomes can affect study results.
Of the 44 consultation cases initiated by trainees, 9 were terminated prior to case completion. In 23 of the remaining 35 cases, data were collected in a way that allowed calculation of effect sizes. The average effect size for these cases was .95. If an effect size of 0 was coded for each of the 9 early terminations, and these cases included in the determination of average effect size, the average effect size would drop to .68.

There are good arguments for using both types of analyses, particularly in studies where both a treatment group and a comparison or control group are used (McCall & Green, 2004; Moher et al., 2010). However, the important point is that in reporting consultation research, the number of participants at each study phase and the type of analysis used in reporting treatment effects should be documented. When individual studies vary in the proportion of participants they use in reporting effect sizes, it is likely that some of the variance in consultation outcomes across studies will be a result of whose data have been included in the determination of treatment outcomes rather than any true variation in consultation outcome.

**Participant Characteristics**

All four methodological sources (i.e., Gresham & Noell, 1993; Task Force on EBI, 2008; Weisz et al., 2005; WWC, 2012) have delineated key variables that are important in describing participants in reports of treatment studies. For school consultation research, there are typically three distinct groups of participants: clients, consultees, and consultants. Issues in describing each group of participants are discussed below.

**Clients.** A description of the client sample, including key demographic variables, is an important component of any description of consultation research. Among the demographic variables most frequently mentioned across the methodological sources are age, gender, ethnicity, family income or socioeconomic status, and grade level in school.

These client characteristics would seem to pose few measurement challenges for the consultation researcher. Despite their relative ease of measurement, it appears that even this basic information about client characteristics is not always included in treatment research reports. In their review of the youth psychotherapy research literature, Weisz et al. (2005) expressed concern that 60% of the psychotherapy studies they reviewed did not provide information on client ethnicity, and 70% did not provide information on family income or SES. A similar situation appears to exist within the consultation treatment literature. In their meta-analysis of 35 school consultation outcome studies, Reddy et al. (2000) reported that only ten provided information on client ethnicity. Given the increasing diversity of the U.S. school population (Federal Interagency Forum on Child and Family Statistics, 2012), information about the extent to which any school-based intervention has been tested and found to be effective with a diverse student body is an important component of establishing its efficacy.

In developing the *EBI Procedural and Coding Manual*, the Task Force on EBI (2008) aimed to incorporate a finer-grained approach to assessing and describing participants than typically had been used in treatment research (when this information is reported at all). They reasoned that more detailed information would permit better assessments of the extent to which an intervention generalized across diverse

The manual calls for intervention and control samples to be described on several dimensions of diversity, including ethnic identity, acculturation, and primary language. It is unclear how some of these dimensions could be efficiently assessed for each participant in treatment research (e.g., ethnic identity and acculturation), although a general characterization of the treatment sample on these dimensions would be helpful in assessing to whom an intervention is likely to generalize. A somewhat different approach to describing participant characteristics is used by the WWC. It has specific protocols for summarizing and evaluating studies in different topical areas, and these protocols describe key participant characteristics that should be reported to describe the study sample. For example, among the characteristics included in the WWC Evidence Review Protocol for Interventions for English Language Learners (2009) are the participants’ native language, English proficiency at pretest, and academic achievement at pretest.

As the Task Force on EBI and others grapple with providing researchers with more sophisticated and sensitive ways of defining diversity and assessing the generalizability of interventions across different contexts and participant populations (Kratochwill & Stoiber, 2002; Miranda et al., 2005), measurement options for describing participants that move beyond gross categorizations by ethnic group and socioeconomic status are likely to become more widely available. In the meantime, it is important to report basic demographic information about clients in consultation research studies.

If the research population of interest is students with a particular clinical diagnosis, another important issue is verification that the research sample met the criteria for the disorder or displayed the symptoms characteristic of the disorder to a significant degree. This information is important in assuring the external validity of an intervention study’s findings. Three out of the four methodological sources discussed issues related to the assessment of participants’ clinical status. In its Reporting Guide for Study Authors (2012), the WWC calls for including a delineation of the eligibility criteria for selecting participants that are members of special populations, as well as their proportion in the study sample. In its listing of external validity indicators, the EBI Procedural and Coding Manual (Task Force on EBI, 2008) asks reviewers to judge the extent to which participants are described in a way that permits other researchers to determine an intervention’s generalizability to its intended participants and also the extent to which the study inclusion criteria are related to the goal of the intervention. For example, if a study examines consultation as a treatment for school phobia, then some participant inclusion criteria should be related to documenting the presence of this disorder or its symptoms.

In their examination of the youth psychotherapy outcome research, Weisz et al. (2005) assessed youth psychotherapy studies on a number of dimensions related to the documentation of client’s initial clinical status. These included whether studies used formal diagnoses based on established diagnostic criteria for various disorders, used a formal clinical cut-off on a standardized clinical measure, or used standardized continuous measures of psychopathology. They found that more than half of the studies of youth psychotherapy in their review identified study participants based on nonstandard measures completed by parents or teachers, or through advertisements or requests for youth with particular problems. When
formal diagnoses or clinical cutoffs on standardized measures were used, which was less than 40% of the time, they were typically based on the use of nonstandard measures with unknown validity.

In their consultation meta-analysis, Reddy et al. (2000) reported that only 57% of consultation studies provided diagnostic or classification information for their client samples. The authors did not report on the types of measures used in these studies and whether standardized instruments with known reliability and validity were used.

When clinical populations are of interest, it is important for consultation researchers to carefully consider how the clinical status of clients is assessed, and whenever possible, use published, standardized measures of the disorder. The data reported by Weisz et al. (2005) and Reddy et al. (2000) suggest adequate documentation of the extent to which study samples represent the clinical populations may be a problem in both psychotherapy and consultation research. However, the current consultation literature also has good models of studies where a clinical population was the focus and participants’ status was established through published, standardized measures (e.g., DuPaul et al., 2006; Fabiano et al., 2010). For example, in a study examining the use of school-based consultation to improve outcomes for students with Attention Deficit Hyperactivity Disorder, Fabiano et al. used published parent and teacher rating scales to evaluate potential study participants, as well as a semi-structured interview with parents. They then asked doctoral-level clinicians to independently review the diagnostic data for each child entering the study and confirm the diagnosis.

Studies examining consultation as a treatment alternative for students identified for special education face a unique difficulty with regard to verifying that study samples meet special education criteria for the classification in question (e.g., learning disability, behavioral/emotional disturbance). Federal special education law provides a number of specific disability categories, but the identification criteria for the categories vary by state, and sometimes within states (WWC, 2012). Furthermore, the manner in which identification criteria are implemented can vary from community to community (Singer, Palfrey, Butler, & Walker, 1989). Consultation researchers investigating treatments for students in special education should provide a clear statement of the classification criteria employed and documentation that their sample met these criteria, as well as a description of the sample on key dimensions related to the disability category in question (e.g., IQ and adaptive behavior for students with mental retardation), using published (versus experimenter-designed) scales. Without this information, the extent to which a sample is comparable to the special education population in different communities and states will remain in question.

In sum, the current state of school consultation research appears to mirror the current state of youth treatment research. In both areas, client descriptive information is sometimes scant, with some studies providing insufficient information about the demographic and clinical characteristics of study participants. Current reviewing standards (e.g., the EBI Procedural and Coding Manual, Task Force on EBI, 2008; WWC, 2009, 2012) and the need to assure that interventions are valid across cultural groups (Miranda et al., 2005) are likely to create a press for much more detailed information about client characteristics in reports of treatment outcome studies.
Consultees. Three of the four methodological sources mentioned the provision of relevant information about treatment agents as an important component of complete intervention outcome reports. The most relevant of these sources to the discussion of consultee characteristics is Gresham and Noell’s (1993) listing of variables likely to moderate consultation treatment outcomes. In addition to basic demographic characteristics (e.g., gender; ethnicity; grade level taught, if teachers are consultees), they listed level of training, experience, classroom management style, attitudes toward consultation, knowledge of classroom interventions, and referral rates for special education and consultation as potential moderators.

Although there is considerable discussion in the consultation literature of the potential impact that consultee characteristics may have on outcomes (e.g., Brown, Pryzwansky, & Schulte, 2011; Gibbs, 1980; Gresham & Noell, 1993), consultation studies often fail to report even basic demographic information on consultees (Reddy et al., 2000). Unfortunately, the lack of routine reporting of basic information about consultees suggests that any systematic variation in outcomes across consultation studies due to consultee characteristics will be less readily detectable in future meta-analyses. A recent study of conjoint behavioral consultation (Sheridan et al., 2012) found that consultation’s effectiveness in changing child behavior was mediated by the improved relationship between the parent and teacher consultee dyad. This finding reinforces the need to fully describe not just clients, but also consultees, to assure the detection of potential mediating or moderating effects operative in consultation.

Consultants. Consultants are the final group of participants to be considered in consultation research studies. As stated in the section discussing consultee characteristics, three of the four methodological sources call for the reporting of basic demographic characteristics of treatment agents (e.g., number used in the study, age, ethnicity, gender), as well as other characteristics that might be potential moderators of treatment outcome or be important in assessing the generalizability of the intervention, such as consultant vocation, experience, training, and education.

Weisz et al. (2005) found that over a quarter of all youth psychotherapy studies failed to report who had provided the intervention. In their review of consultation outcome research, Reddy et al. (2000) found that less than 10% of studies reported data on consultant ethnicity, 40% on consultant gender, and 60% on their educational level. It appears that in both the child psychotherapy and consultation research areas, descriptions of treatment providers are quite limited. Recent research in consultation varies considerably in its reporting of consultant characteristics. For example, no information beyond that the consultants were school psychology graduate students was reported in Fabiano et al. (2010). However, more detailed information was provided by Sheridan, Warnes, Woods, Blevins, Magee, and Ellis (2009) and DuPaul et al. (2006).

Summary
Characteristics of the client, consultee, and consultant are likely to affect consultation outcomes (Gresham & Noell, 1993). Given the limited information available about consultation participants, the conclusions of Weisz et al. (2005)
concerning the state of youth psychotherapy research appear to also apply to school consultation research:

Our review suggests that the critical first step toward moderator assessment—i.e., collecting information on participant characteristics that might moderate effects—has not been taken in a remarkably large percentage of studies . . . A related problem noted in our review is the high percentage of studies in which no data were reported for important variables such as who carried out the treatment and in what setting the treatment took place. Given the potential value of all these types of information to the field, and the lost opportunities for moderator assessment once data sets are no longer available, it seems desirable to encourage greater consistency in the kinds of information required by journals prior to acceptance of manuscripts for publication. (p. 358)

The evolving standards for reporting treatment research call for more detailed descriptions of study participants and settings than has been true in the past (e.g., APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Weisz et al., 2005). At present, the most salient problem in describing school consultation participants is not a lack of available measures, but failure to include descriptive information even on easily measured variables. However, as school consultation research moves beyond basic questions about potential moderator variables, there will be a need for more refined ways of describing participants on a wider range of dimensions. One important direction for future research will be the development of measures that allow researchers to ask more sophisticated questions about moderators, once they are identified. For example, it may be possible to examine whether consultation treatments are differentially effective by client ethnic group, and whether this difference is explained by cultural, language, background experience, or socioeconomic factors.

More complete reporting of characteristics of all participants will also allow an assessment of how representative the samples in consultation research are to the treatment providers and children served in the schools. For example, in a review of youth psychotherapy studies, Weisz and Gray (2008) reported that less than 2% of youth psychotherapy studies included any practicing clinicians. It would be useful to have similar information about the representativeness of treatment providers in consultation studies compared to those who provide those services in the schools.

Measurement of Consultation Process

Consultation process is used here to refer to actions of the consultant or consultee that have an impact (or are thought to have an impact) on the outcome of consultation. It is the “what” and “how” of consultation. The primary focus of this section will be measurement issues related to describing what constituted the treatment in a consultation study, and assessing the extent to which that treatment was implemented as planned (treatment integrity). Readers are referred to the chapters by Noell and Gansle, as well as by Erchul, Grissom, Getty, and Bennett, in this volume for discussion of topics of interest to consultation researchers within the area of consultation process.
Describing and Documenting Treatments

For many years, methodologists have noted that treatment outcome studies often lack a detailed description of the actual treatment provided, and/or fail to provide evidence that the treatment described was the treatment that was delivered (e.g., Gresham & Kendell, 1987; Hagermoser Sanetti, Gritter, & Dobey, 2011; O’Donnell, 2008; Sechrest, West, Phillips, Redner, & Yeaton, 1979). Yet, a full description of a treatment, evidence that it was implemented as intended, and information about the extent to which it differed from any alternatives it was tested against, are central to making correct causal inferences about a treatment, as well as replicating or disseminating the treatment should the outcomes be positive. In other words, treatment integrity (also called treatment fidelity), or the specification and measurement of the independent variable, is fundamental to the goals of treatment research (Cordray & Pion, 2006).

Although all four methodological sources mention the importance of adequately specifying the independent variable in treatment studies (i.e., Gresham & Noell, 1993; Task Force on EBI, 2008; Weisz et al., 2005; WWC, 2012), only two of the sources explicitly discuss procedures for describing and documenting psychological treatments (i.e., Weisz et al., 2005; Task Force on EBI, 2008). Both of these sources put considerable emphasis on the use of treatment manuals. This emphasis is not surprising as treatment manuals have come to be the predominant means of specifying interventions (Miller & Binder, 2002).

Although treatment manuals make replication of psychological interventions much easier, their use does not fully address the concerns raised about the specification of the independent variable in treatment research (Cordray & Pion, 2006). Without direct measurement of the extent to which the treatment providers adhered to the procedures described in the manual, we have no evidence that the intervention described was the intervention implemented and no way to assess variability in how a treatment was implemented across participants (Hagermoser Sanetti et al., 2011; Sass, Twohig, & Davies, 2004). Thus, adequate assessment of treatment integrity requires measures of the extent to which the providers adhered to the treatment protocol. Others have also suggested that competence in delivering the intervention should also be assessed, as well as several other potential dimensions of treatment integrity, including dosage (i.e., number of times the treatment was delivered) and treatment differentiation (Durlak & DuPre, 2008; Perepletchikova, 2011; Schulte, Easton, & Parker, 2009; Waltz, Addis, Koerner, & Jacobson, 1993).

Added to the usual challenges in assessing treatment integrity for educational and psychological treatments, consultation researchers face the problems posed by the two-tiered nature of the consultative services. The consultant interacts with the consultee, who then provides the treatment to the client. As such, school consultation research must not only provide information about the model of consultative problem solving and how it was implemented (typically by describing interactions between the consultant and consultee), but also information about how the consultee implemented the planned intervention with the client.

Noell and Gansle (this volume) provide a thorough discussion of issues of treatment integrity in school consultation research in the context of this two-tiered model. Reflecting the two-tiered nature of consultation processes, they use the term
**Measurement in School Consultation Research**

consultation procedural integrity to refer to the extent to which consultation procedures were implemented as designed, and the term intervention plan implementation to refer to the integrity with which an intervention plan developed in consultation was implemented. In this section, issues related to the measurement of these two aspects of treatment integrity will be discussed separately.

**Consultation Procedural Integrity**

Although treatment manuals are only a first step in assuring that an intervention is adequately described (Cordray & Pion, 2006), even this first step is not yet a universal in treatment research. Despite considerable emphasis on treatment manuals in the youth psychotherapy literature, Weisz et al. (2005) found only about half of youth psychotherapy studies used treatment manuals, although an additional one-third used a structured treatment protocol, consisting of a detailed listing of the steps involved in the intervention.

The percent of consultation outcome studies using manuals is unknown, as their use has not been tracked in any comprehensive review of school consultation. However, a number of consultation treatments within the behavioral model have been “manualized,” providing specific procedures that should be followed by the consultant to implement consultative problem solving. For example, Bergan’s (1977) initial book, Behavioral Consultation, and its successor, Behavioral Consultation and Therapy (Bergan & Kratochwill, 1990), each provide a detailed set of instructions that the consultant is to follow in implementing consultative problem solving. In addition, Sheridan and Kratochwill (2008) have published a manual for Conjoint Behavioral Consultation, a model of behavioral consultation that has been expanded to include professionals (typically teachers) and parents as joint consultees. D. Fuchs and his colleagues (Fuchs et al., 1989) also have published a manual for their application of Bergan’s (1977) model to prereferral assessment. Finally, the Consultant Evaluation Rating Form (CERF) Scoring Manual (Hughes & Hasbrouck, 1997) for Hughes and colleagues’ Responsive Systems Consultation model (Denton, Hasbrouck, & Sekaquaptewa, 2003; Hughes, Hasbrouck, Serdahl, Heidgerken, & McHaney, 2001) is sufficiently detailed that it could serve as both a treatment manual and procedural integrity tool for implementing this model of consultation.

**Consultant adherence to consultation procedures.** To the extent that treatment manuals constitute an explicit statement of how the treatment process should proceed, they then provide the basis for developing measures of adherence and competence in implementing the model (Waltz et al., 1993). Although adherence can be assessed in a variety of ways, including self report, case notes, and objective measures (Noell and Gansle, this volume), consultation researchers appear to have most frequently assessed consultation procedural integrity by having independent judges compare a list of objectives for each stage of consultative problem solving to audiotapes of consultation sessions. Typically, the objectives have been taken (or created) from a treatment manual, and the number of objectives achieved has been divided by the total number of objectives stated, converted to a percentage, and used as a measure of consultation procedural integrity. Equal weight is given to each objective.
For example, in Sheridan et al.’s (2012) RCT comparing the conjoint behavioral consultation model to schools’ usual methods for dealing with students with disruptive behavior (e.g., office referrals, student assistance teams), independent coders assessed consultants’ adherence to interview objectives using checklists for each interview while listening to audiotapes of the interviews. The authors reported adherence in terms of a percent of interview objectives met (average of 98% or higher for each interview type), with an overall interrater agreement of 96%. DuPaul et al. (2006) used checklists completed by independent observers to examine the extent to which consultants followed their behavioral consultation model’s specified procedures in working with teachers to address the academic needs of students with attention-deficit/hyperactivity disorder.

Within a study, the number of sessions coded to establish the percent of consultation objectives achieved has ranged from about 20% (e.g., DuPaul et al., 2006) to 100% (e.g., Sterling-Turner, Watson, & Moore, 2002). When less than 100% of interviews are coded, there is presently no empirical basis for judging the adequacy of the sampling strategy, although Perepletchikova (2011) suggested a random sample of 20–40% of treatment sessions be used to assess treatment integrity in a discussion of methodological guidelines for assessing treatment integrity.

Only a few consultation studies have reported the criteria they used for judging whether adherence to interview objectives was adequate (e.g., Kratochwill et al., 1995), and in some cases, the results of the adherence checks have not been reported, only that the adherence checks were done (e.g., Jones, Wickstrom, & Friman, 1997). However, when adherence has been reported in studies, it has generally been in the 85 to 100% range.

At present, the use of checklists with a simple calculation of the percent of treatment objectives achieved appears to be the most widely used practice in assessing consultant adherence as it relates to consultation procedural integrity. Advantages of this method are that it is objective and its reliability can be determined. However, as the technology for assessing procedural integrity advances, it may be useful to consider whether some objectives should be weighted more heavily than others, and to develop an empirical basis for determining what constitutes adequate model adherence (Mowbray, Holter, Teague, & Bybee, 2003; Sheridan, Swanger-Gagne, Welch, Kwon, & Garbacz, 2009). In terms of weighting some objectives more heavily, most persons would view obtaining a behavioral description of the client behavior in a problem analysis interview as more critical than setting a date for the next interview. However, because both are objectives listed for the initial interview in behavioral consultation, they have sometimes been given equal weight in assessing model adherence (e.g., Bergan & Kratochwill, 1990; DuPaul et al., 2006).

**Going beyond adherence in assessing consultation procedural integrity.** In a thoughtful discussion of the measurement of treatment integrity, Waltz et al. (1993) suggested that measures of treatment adherence would be improved by using manuals that specify all aspects of treatment considered key to its successful implementation, including nonspecific aspects such as warmth and nurturance. They also suggested that adherence be assessed relative to treatment agent actions that are: (a) unique to a treatment modality; (b) essential, but not unique; (c) compatible, but
Measurement in School Consultation Research

neither necessary nor unique; and (d) proscribed. As an example, they suggested that assigning homework would be essential and unique to behavioral therapy, but proscribed in psychodynamic therapy. In contrast, empathetic listening would be essential to both modalities, but not unique.

A more comprehensive approach to assessing treatment integrity that includes assessment of prescribed and proscribed behaviors has since been used in some treatment studies, particularly in studies of psychotherapy (e.g., Barber et al., 2006). Several models of treatment integrity have been proposed that broaden the construct well beyond simple percentages of treatment steps implemented (e.g., Bellg et al., 2004; Durlak & DuPre, 2008; Perepletchikova, 2011; Power, Blom-Hoffman, Clarke, Riley-Tillman, Kelleher, & Manz, 2005). For example, instruments are available that can be used to characterize therapist behaviors across multiple psychotherapy models (e.g., McLeod & Weisz, 2010). This broader approach has proved useful in quantifying the extent to which treatment was delivered in treatment studies, as well as examining the extent to which treatment and comparison conditions truly differed, or how treatment agent actions changed over the course of treatment (Weersing, Weisz, & Donenberg, 2002).

Developing tools to assess consultation process integrity more broadly, as has been done in psychotherapy, could prove fruitful in a number of ways. First, articulating and assessing consultant behaviors in the four categories originally proposed by Waltz et al. (1993) would provide an explicit basis for characterizing consultation models in terms of commonalities and differences. Second, such a listing would provide a basis for developing consultation process integrity measures that assess specific and nonspecific aspects of consultant behavior, and allow for assessment of errors of inclusion or exclusion by consultants in treatment studies. A detailed articulation of consultation process would also help in examining overlap between models of consultation in treatment comparisons, operationalizing ill-defined constructs such as “collaboration” that may have many different meanings across consultation models and researchers (Schulte & Osborne, 2003), and specifying the “active ingredients” in consultation that researchers hypothesize account for its effects (Sheridan, Rispoli, & Holmes, 2014).

Another important issue relevant to consultation procedural integrity is treatment agent competence (sometimes termed “quality” of treatment delivery). Contemporary models of treatment integrity (e.g., Durlak & DuPre, 2008; O’Donnell, 2008; Perepletchikova, 2011; Power et al., 2005) treat adherence and competence as separate dimensions of treatment integrity, and empirical investigations of these dimensions in psychotherapy have provided some support for this proposition (see Barber, Sharpless, Klostermann, & McCarthy, 2007, for a review of this literature). Competence is defined in terms of the extent to which the treatment agent took relevant aspects of the therapeutic context into account and responded to these appropriately, including knowing when and when not to intervene (Barber et al., 2007; Waltz et al., 1993). Taking into account relevant contextual factors in consultation might include adapting procedures to the: (a) knowledge and skill level of the consultee, (b) degree of consultee distress, or (c) particular client problem. For example, proceeding lockstep through a problem identification interview while a stressed consultee cries in frustration over her difficulties managing a client would...
constitute an example of adequate adherence to behavioral consultation, but consultant incompetence. Alternately, a consultant and consultee might achieve all objectives related to developing an intervention plan in a consultation interview (high adherence), yet develop a plan that is very unlikely to succeed in changing client behavior (low competence, see Fuchs & Fuchs, 1989, 1992).

Although procedures for assessing consultant competence are less developed than those for assessing adherence, there are some studies that have examined consultant competence. For example, Kratochwill et al. (1995) used tests of knowledge of learning theory and behavior consultation to assess consultant trainees’ competence prior to their first consultation case. However, contextually sensitive deployment of behavioral knowledge during actual consultation cases was not assessed. In an early attempt to assess consultant competence in a way that did take consultation context into account, Bergan and Tombari (1976) examined the variety of psychological principles applied by consultants across consultation cases, hypothesizing that the consultant who applied a broad range of psychological principles would be more effective than a consultant who only used a narrow range of principles. Working from case reporting forms filled out by consultants, they coded what change procedures were incorporated into client intervention plans developed by consultant and consultees (e.g., modeling, positive reinforcement, task alteration). An index of consultant flexibility was then calculated for each consultant based on the number of different principles used in plans and the proportions of cases in which each psychological principle was employed. They found that consultant flexibility was a predictor of successful problem resolution. Although Bergan and Tombari’s procedures did not measure contextually sensitive deployment of consultant strategies directly, presumably the consultant who used a range of psychological principles would be more able to respond to the unique aspects of the client’s difficulty.

Hughes et al. (2001) used the Consultant Evaluation Rating Form (Hughes & Hasbrouck, 1997) to evaluate consultants’ implementation of the Responsive Systems Consultation model. This measure explicitly defines and measures consultant competence on both task-oriented and interpersonal dimensions of consultation process, including effectiveness of the intervention plan, consultant nonverbal behavior, and consultant sensitivity to consultee needs. The reliability and validity data for the measure are promising (Hughes et al., 2001). Within consultation, this coding system appears to come the closest to the recommendations for developing treatment agent adherence and competence measures recommended by Waltz et al. (1993) and Perepletchikova (2011). Sheridan et al.’s (2014) broadened framework for assessing treatment integrity in conjoint behavioral consultation, and the instruments based on it, also go well beyond a “percent of objectives achieved” approach to adherence and allow an examination of how effectively key aspects of consultation were implemented.

**Intervention Plan Implementation**

The second component of treatment integrity in consultation is implementation of the intervention plan by the consultee (Noell and Gansle, this volume). In most models of consultation, providing the intervention is the responsibility of the consultee. Given that researchers exercise less control over the behavior of consultees
than consultants in studies, intervention plan implementation is more likely to vary among cases. As such, measurement of intervention plan implementation is particularly important in documenting that consultative problem solving is responsible for client change. Measurement issues associated with assessing the extent to which the consultee adhered to the intervention plan developed in consultation, as well as consideration of broader aspects of treatment integrity, such as consultee competence in implementing the intervention, are addressed below.

**Consultee adherence to the intervention plan.** Intervention plan adherence has not consistently been assessed in consultation research (Gresham, 1989; Gresham & Kendall, 1987; Noell and Gansle, this volume; Sheridan et al., 1996; Sheridan et al., 2009). However, when assessed, it has been measured in a variety of ways, including consultee self reports (e.g., Evans, Schultz, & Serpell, 2008), permanent products (e.g., Fabiano et al., 2010; Noell et al., 2005), and direct observations (e.g., Sheridan et al., 2012).

Not surprisingly, self reports of intervention plan adherence are viewed as problematic (Mowbray et al., 2003). However, consultee accuracy may depend on the form of self report. For instance, in a study consisting of a series of three case studies in consultation, Robbins and Gutkin (1994) simply asked consultees if they had implemented the intervention plan developed in consultation. All consultees reported they had implemented the intervention plan; however, observations of the frequency with which one of the central components of the intervention was implemented indicated that implementation frequency was quite low. In contrast, Ehrhardt, Barnett, Lentz, Stollar, and Reifin (1996) developed intervention scripts collaboratively with parent and teacher consultees. The scripts were designed to provide guidance in implementing interventions in consultees’ own words. The scripts were then adapted to a checklist format that allowed consultees to carry out the intervention and monitor their own degree of implementation of the treatment at the same time. Comparisons of consultees’ self ratings and independent observers’ ratings indicated high compliance with the intervention plan and high accuracy in self report. More recently, Hagermoser Sanetti and Kratochwill (2009) have developed a Treatment Integrity Planning Protocol (TIPP) for use in behavioral consultation. This protocol introduces a formalized multi-stage process for planning how to implement a classroom intervention that involves creating a self-assessment of treatment integrity for the consultee. In a test of the TIPP in a small n, multiple baseline study, Hagermoser Sanetti and Kratochwill reported quite high levels of agreement between three consultees’ self reports of the number of intervention steps they implemented each day and permanent product evidence of their implementation of the same intervention steps.

Although some forms of self report may be accurate, they are probably best viewed as a supplement to an independent measure of intervention plan adherence rather than the sole measure of intervention plan adherence in treatment research. However, when a self-monitoring tool serves as a reminder of the correct implementation sequence for the consultee, it may be useful both as an assessment tool and as a means of promoting intervention plan adherence (Hagermoser Sanetti, Chafouleas, Christ, & Gritter, 2009; Plavnik, Ferreri, & Maupin, 2010).
Permanent product assessment of intervention plan adherence in consultation was suggested by Gresham (1989) and has been used extensively by Noell in his studies investigating variables that affect consultee intervention plan implementation (Noell, Witt, Gilbertson, Ranier, & Freeland, 1997; Noell et al., 2000; Noell et al., 2005), as well as by others (Fabiano et al., 2010; Sheridan et al., 2009). With this technique, the intervention designed in consultation is broken down into smaller steps such that implementation of each step results in the generation of a permanent product. For example, daily ratings of student behavior on a home-school report card might be used as an indicator that a teacher had consistently evaluated a student’s behavior, and the parent’s signature on the report card would indicate that the report card had been brought home and the agreed-upon consequence for a particular rating implemented by the parent.

Three advantages of this measurement method are that it is likely to be less reactive than observations (Noell and Gansle, this volume), is low cost, and does not require sampling because a measure of intervention plan implementation is generated each time the intervention is used. However, the use of permanent products does not lend itself to the assessment of implementation for all types of interventions. For example, it would be difficult to envision how a permanent product assessment could be generated for an intervention plan in which a teacher ignores inappropriate behavior, particularly if the inappropriate behavior has had a high frequency.

One potential advantage of the increasing use of computers and the internet in the home and classroom is that permanent-product measures that do not require extra time or effort from the consultee can be incorporated into the intervention. For example, the online reading instructional system Headsprout (www.headsprout.com/) generates online performance reports that could be used to indicate the extent to which reading instruction in a specific area was carried out. As computerized and internet-based interventions become more sophisticated and inexpensive, their use in plans developed in consultation may increase because of the ease with which intervention plan implementation can be documented with them.

Direct observation of intervention plan implementation, the final method of assessing plan implementation, is adaptable to a wide range of interventions. It is considered the “gold standard” of integrity assessment (McLeod & Islam, 2011). However, unless the intervention is only delivered on an infrequent basis, it is likely to be an expensive and intrusive method of monitoring plan adherence. One option for reducing the cost of observation is sampling only a proportion of the times in which the intervention is to be implemented. For instance, Sterling-Turner et al. (2002) assessed consultees’ adherence with intervention plans that involved ignoring inappropriate behavior, administering reinforcers, and providing prompts with classroom observations lasting 20 to 40 minutes per day. However, unless the intervention is only implemented for that brief period each day, the data collected with these types of sampling plans reflect only a small portion of the school day. The generalizability of the results obtained during the observation period to the entire day, or to all instances where the intervention plan is to be implemented, is not known.

Although consultee self report, permanent products, and direct observation have been presented separately here as means of documenting intervention plan adherence, they can also be combined. Most contemporary discussions of treatment integrity acknowledge the need for multi-method approaches to treatment integrity.
to assure that as many critical aspects of an intervention as possible are measured (e.g., Hagermoser Sanetti & Kratochwill, 2008; Resnick et al., 2005; Sheridan et al., 2009). For example, in developing their treatment integrity data collection plan, Sheridan et al. (2012) cited the presence of multiple component interventions, delivered in a contingent manner in consultation, as a primary justification for their use of multiple methods in assessing treatment integrity in a study of conjoint behavioral consultation. They reported that only about half of the critical components of their home and school interventions could be captured with permanent product measures, and therefore they supplemented permanent product measures with both self report and direct observation measures of intervention implementation.

Going beyond adherence in measuring intervention plan integrity. Despite the fact that consultation researchers still face many challenges in developing consultee intervention plan adherence measures, we do have examples of these measures in the consultation literature (e.g., Noell et al., 2005; Sheridan et al., 2012). However, measures of other aspects of treatment integrity, such as consultee competence in implementing the intervention plan, are much more difficult to find. Again, the distinction made in the psychotherapy literature between therapist adherence and competence is between (a) accomplishing specified objectives or exhibiting certain frequencies of desired behaviors (adherence), and (b) examining the adequacy with which these objectives were accomplished or how sensitively particular behaviors were deployed given the context (competence). For consultation, it is conceivable that consultees might adhere to intervention plans but show low competence. For example, a consultee might use contingent praise so frequently that he or she disrupts instruction or draws unwanted attention to a child, or might fail to adapt an instructional intervention when a child is unable to master a critical prerequisite to further instruction and clearly is not comprehending the instruction provided.

It may be that adequate specification of all the components of an intervention would result in complete overlap between a measure of consultee adherence and competence. This situation would simplify the measurement of both constructs. Given that most assessments of treatment agent competence involve the use of expert raters (Kazantzis, 2003), one potential test that a intervention plan adherence measure also assesses consultee competence might be a high correlation of the intervention adherence measure with expert ratings of videotapes of the consultee implementing the intervention.

Another key component of treatment integrity that has not received sufficient attention in consultation is dosage (Sheridan et al., 2009). In consultation intervention plan integrity, dosage corresponds to the number of times the intervention is implemented. For example, consider a intervention plan to use the “cover-copy-compare” strategy to master subtraction facts (Codding, Hilt-Panahon, Panahon, & Benson, 2009). A teacher who implements all components of this intervention with integrity, but only delivers the intervention for a total of four 10-minute sessions across a period of two weeks, is less likely to have an impact on a student’s mastery of subtraction facts than a teacher who implements the same intervention for 10 minutes a day, five days a week, for a month.

Intervention plan implementation measurement technology is still in its infancy. Although studies are beginning to assess intervention plan implementation,
we have relatively little data concerning the reliability and validity of measurement procedures in this domain. Specifically, validating a particular intervention plan implementation measure will require documentation of the extent to which the measure: (a) represents the intervention (content validity); (b) generalizes across time intervals, situations, and settings where the intervention is used; (c) converges with other measures of intervention implementation; and (d) discriminates between adequate and inadequate adherence (Gresham, 1989; Mowbray et al., 2003). Although we have some initial findings in this area for the measurement of intervention plan implementation in consultation (Sheridan et al., 2009), as more consultation researchers develop intervention implementation measures and provide data about their psychometric characteristics, this technology should develop considerably.

**Summary**

Treatment integrity in consultation can be divided into two distinct components: consultation procedural integrity and intervention plan implementation. Each of these components can be considered along multiple dimensions, such as dosage, adherence, and competence. At present, there are good examples in which treatment integrity has been assessed in consultation research, but it has not been routinely done. This situation parallels the state of research in youth psychotherapy research (Weisz et al., 2005). For both components of treatment integrity in consultation, adherence has been assessed much more frequently than competence.

The development of treatment integrity measures for consultation is still in its beginning stages. There remain many unanswered questions about the frequency and form of measurement needed for reliable and valid assessments of consultation procedural integrity and intervention plan implementation. Consultation researchers should consider the comprehensive framework proposed by Waltz et al. (1993) in developing measures of treatment integrity, particularly measures of consultation procedural integrity. This framework includes specification of behaviors that are (a) unique to a treatment modality; (b) essential, but not unique; (c) compatible, but neither necessary nor unique; and (d) proscribed. When employed in the consultation context, such a framework would result in clearer descriptions of models; allow delineation of overlap between models; and promote specification of frequently studied, but poorly defined, variables in consultation, such as collaboration.

**Measurement of Client Outcomes**

In any type of intervention outcome research, accurately assessing the impact of intervention is a critical issue. Three aspects of outcome assessment are considered in this section: (a) what should be measured in assessing client outcomes in consultation, (b) how it should be measured, and (c) methods for measuring consultation outcomes when studies address a diverse set of referral concerns.

**What to Measure**

The four methodological sources (i.e., Gresham & Noell, 1993; Task Force on EBI, 2008; Weisz et al., 2005; WWC, 2011, 2012) discuss a number of potential areas that are relevant to evaluating the impact of consultation. All sources mention the importance of assessing the problem specifically targeted for treatment, whether
it is a clinical condition or specific target behavior. Additional areas of outcome assessment are: (a) client symptoms outside the primary focus of treatment, (b) functional impact of the treatment (e.g., improved grades, fewer school suspensions, change in educational placement), (c) consumer satisfaction (or social validation of treatment effects), (d) environmental impact (e.g., reduced parenting stress or increased instructional time when a student’s disruptive behavior is controlled), and (e) adverse effects of treatment. Also discussed are generalization of behavior changes over settings and maintenance of behavior change after the cessation of treatment (Task Force on EBI, 2008).

Weisz et al. (2005) reported that the typical psychotherapy outcome study employed a total of 12 participant measures. All studies had at least one measure of the target problem, with the average study including five. Seventy-eight percent of the studies included a measure of symptoms outside the primary focus of treatment and 28% included one or more measures of the functional impact of treatment. Less than 10% of the studies included a measure of consumer satisfaction or environmental impact of treatment. Weisz and colleagues did not tabulate studies’ use of measures to assess adverse effects of treatment or generalization and long-term maintenance of gains, although Weisz, Weiss, Han, Granger, and Morton (1995) reported that a third of youth psychotherapy studies included follow-up measures.

There have been no examinations of the breadth of outcome areas assessed in consultation research in the detailed manner of Weisz et al. (2005). Gresham and Noell (1993) characterized consultation research as employing dependent variables that were “limited in scope, univariate, and ecologically invalid” (p. 257). However, this characterization no longer seems valid. Most contemporary consultation studies use multiple measures and assess many of the recommended domains.

For example, Sheridan et al.’s (2012) study of conjoint behavioral consultation included measures of client disruptive behavior across school and home, measures of client’s adaptive and social skills (all in the targets of treatment domain), a measure of treatment acceptability (consumer satisfaction domain), and a measure of the parent-teacher relationship (environmental impact domain, although the parent-teacher relationship was also examined as a mediator).

DuPaul, Jitendra, and colleagues (DuPaul et al., 2006; Jitendra et al., 2007) assessed client achievement and client academic behavior (targets of treatment domain), and report card grades (functional outcomes domain), as well as teachers’ ratings of acceptability of the interventions (consumer satisfaction domain) in their study of consultation as a means to enhance school functioning for students with ADHD. Fabiano et al. (2010) included observations of clients’ classroom disruptive behavior, and ratings of students’ ADHD symptoms and classroom academic functioning (all in the targets of treatment domain). They also collected ratings of students’ improvement on individualized education plan goals (functional outcome domain), teacher and parent satisfaction in their study of school consultation (consumer satisfaction domain), and teacher-student relationship (client symptoms outside the primary area of treatment domain). In one of the more interesting measures in the functional outcomes domain, Schultz, Evans, and Serpell (2009) reported that consultation delayed the onset of academic failure (i.e., earning a grade point average below 1.0) for middle school students compared to students who did not receive consultative services.
How to Measure

Across the four methodological sources (i.e., Gresham & Noell, 1993; Task Force on EBI, 2008; Weisz et al., 2005; WWC, 2011, 2012), recommended practices for assessing outcomes include the use of measures that: (a) are objective, (b) are reliable and valid, (c) include assessment of normalization, and (d) allow assessment of each treatment construct with a multi-method, multi-source approach. These four characteristics of high quality outcome measurement are discussed below. In addition, the issue of how to measure outcomes across referral problems, a particular problem in consultation research, is addressed.

Objective measurement. The strongest evidence that a treatment is effective is produced when: (a) assessors are blind to the participant’s treatment status; (b) clients are unaware of or unable to influence the ratings; or (c) independent life events, such as arrests or high school drop-out rate, are used as outcome measures (Moher et al., 2010; Weisz et al., 2005). When the child or those involved in treatment serve as informants or are able to influence the intervention assessment, the evidence of an intervention effect is weakened because factors other than actual behavior change may have influenced the outcome measure. The preference for objective measurement does not mean that teachers’ and parents’ views of whether treatment has resulted in changes in child symptoms or behavior are unimportant. Rather, these assessments should be viewed as a component of social validation (Carter, 2010; Finn & Sladeček, 2001) or consumer satisfaction rather than as the primary indicator of treatment outcome.

Weisz et al. (2005) reported that approximately 63% of youth psychotherapy studies had included behavioral observations among their outcome measures. Less than half of these, or a third of all studies reviewed, had used “blind” observers to collect the behavioral observations. Approximately 14% of all studies had used independent life data as an outcome measure. Using a somewhat different definition of direct observation, Sheridan et al. (1996) reported that direct observation was used in 44% of consultation outcome studies. They did not report the percent of consultation studies using blind observers or independent life data. Although direct comparisons of the two treatment literatures are not possible, it appears that less than half of all studies of either treatment modality employed objective observers in assessing client improvement.

Although the use of independent observers helps to assure assessments of treatment outcome are unbiased, results from independent observations often are based on very limited observation opportunities, and that may be a disadvantage relative to less objective measures completed by consultees, who may have more opportunities to observe the student. For example, in Fabiano et al.’s (2010) study, their primary outcome measure, frequency of rule violations, was assessed through three 30-minute classroom observations that occurred at the beginning of the study, before active treatment was initiated, and when the treatment was fully implemented.

Given that Lomax (1982) found that several hours of classroom observation are needed to obtain stable estimates of pupil behavior, it is unlikely that the amount of classroom observation used in many consultation studies would yield comparisons between treatments conditions that were sensitive to anything but very large changes
in pupil behavior. This fact does not diminish the importance of objective outcome measures, but suggests that consultation researchers using observational measures should consider carefully the frequency and length of behavioral observations that will be needed to assure that a treatment effect will be detected, or take steps to increase the likelihood that their measures would be sensitive to treatment effects. For example, Fabiano et al. (2010) asked parents to obtain physician permission to withhold ADHD medication from study participants on the days of classroom observations to decrease the probability of low rates of disruptive behavior due to medication effects during the classroom observations. Another alternative is the use of objective measures that are less expensive and time intensive than classroom observations. For example, when academic functioning is the focus of consultation, standardized measures of achievement administered by persons blind to treatment condition are appropriate outcome measures (e.g., DuPaul et al., 2006).

**High reliability and evidence of validity.** One of the eight features of studies examined in evaluating the quality of the evidence to support a treatment in the *EBI Procedural and Coding Manual* (Task Force on EBI, 2008) is the use of outcome measures that produce reliable and valid scores. The highest rating of “strong evidence” in the coding system’s key feature of measurement requires the use of measures for each outcome construct with reliability coefficients above .85 and evidence of validity. Similarly, WWC (2011) excludes intervention studies from review if their outcome measures are not reliable or valid, but fails to specify the criteria for the results in a measure being deemed reliable or valid. However, it appears that published instruments are generally considered reliable and valid, as the guidelines state these need not be examined as closely as unpublished measures.

To date, no review has characterized outcome measures across studies in youth psychotherapy or consultation studies in terms of their reliability and validity. However, it seems reasonable to assert that many of the outcome measures used in school consultation studies would meet the criteria outlined in the *EBI Procedural and Coding Manual* (Task Force on EBI, 2008). That is, most published consultation studies have provided evidence that their outcome measures are reliable across at least one facet of measurement (e.g., observers, items), and relate to the construct assessed (e.g., through construct, content, or concurrent validity). However, it is also true that there is considerable variability in outcome measures across consultation studies, even when the same outcome construct is assessed. For example, student off-task behavior has been assessed through standardized observational systems (Dunson, Hughes, & Jackson, 1994), researcher-designed observational systems (Fabiano et al., 2010), and norm-referenced teacher ratings of attentional behavior (DuPaul et al., 2006). Although assessing the same construct, these measures are likely to vary in a number of ways. This variability is likely to contribute to differences in treatment outcomes and estimates of consultation effect size across studies.

As noted earlier, Wilson and Lipsey (2001) provided evidence that how a treatment outcome is measured contributes as much in the variability in treatment outcomes across studies as what is measured. In addition, two comprehensive reviews have found that intervention effects are likely to be larger when researcher-designed measures rather than published or standardized measures are used (Marshall,
Lockwood, Bradley, Adams, Joy, & Fenton, 2000; Wilson & Lipsey, 2001). Taken together, these findings suggest that moving toward the use of a small number of standardized, published instruments for common targets of consultation treatment (e.g., achievement, on-task behavior) would likely yield more consistent and generalizable estimates of intervention effects for consultation. If the same instruments were used with other interventions aimed at the same problems, this strategy would also allow comparisons among treatment approaches.

More consistent use of the same outcome measures in consultation studies with similar treatment targets would also make psychometric studies of these measures more cost effective. For example, given the earlier discussion of the limited amount of occasions typically sampled when observational measures are used in consultation studies, it would be very useful to know the number and length of observational sessions that would be needed to have a reasonable likelihood of obtaining a generalizable or stable estimate of on- and off-task behavior when planning a consultation study.

Normalization. One way of assessing the educational or clinical significance of a change in client behavior is the extent to which the client’s behavior is normalized by treatment. Both the *EBI Procedural and Coding Manual* (Task Force on EBI, 2008) and Gresham and Noell (1993) discussed the importance of including measures that allow this type of assessment as part of treatment studies.

Neither Weisz et al. (2005) nor any of the comprehensive consultation reviews that have been conducted to date (e.g., Fuchs, Fuchs, Dulan, Roberts, & Fernstrom, 1992; Reddy et al., 2000; Sheridan et al., 1996) have examined the extent to which client normalization has been assessed in youth psychotherapy or consultation outcome studies. However, examples of normalization assessment are present in the consultation literature.

For example, Sheridan, Kratochwill, and Elliott (1990) compared the social initiations of withdrawn children and their nonreferred classroom peers in assessing the impact of behavioral consultation on withdrawn children’s social functioning. This strategy employs local “micronorms” (Gresham & Noell, 1993) as the basis for a normative comparison. Two other strategies for normative comparisons are examining the number of clients who no longer meet diagnostic criteria following treatment, and assessing number of clients who score within normal limits (e.g., within one standard deviation) on a norm referenced measure of the target behavior before and after treatment. The latter strategy was employed by Fabiano et al. (2010) who compared the percent of students in their consultation and comparison groups that fell within one standard deviation of the mean for normative samples in terms of teacher ratings of ADHD and oppositionality at the study endpoint.

Multi-source, multi-method approach. Both the *EBI Procedural and Coding Manual* (Task Force on EBI, 2008) and Gresham and Noell (1993) have advocated the use of multi-source, multi-method approaches to assessing treatment outcomes. The use of multiple measures to establish treatment efficacy presents an interesting methodological dilemma. On the one hand, measuring the same construct through multiple methods and sources helps assure that the treatment effect is robust and not
dependent on the way it is measured (Cook, 2000). On the other hand, this practice increases the Type I error rate unless the researcher adjusts the alpha level required for significance to reflect the use of multiple measures. The increased alpha level, in turn, requires a larger sample size to achieve adequate power. When resources are limited, the researcher may need to sacrifice power to meet the demand for multiple measures.

The WWC Procedures and Standards Handbook (WWC, 2011) has endorsed a specific procedure, the Benjamini-Hochberg (BH) correction (Benjamini & Hochberg, 1995), to adjust significance levels when multiple outcome measures within one domain have been used to assess the impact of an intervention. Variants of the correction are available for two group and multiple group comparison designs. Their procedure is less stringent than the Bonferroni correction. An example of the use of the BH correction in consultation research is found in Sheridan et al. (2012). Other contemporary consultation studies have used multivariate statistical procedures that control for multiple comparisons, such that a correction procedure is not needed (e.g., Jitendra et al., 2007).

Measuring consultation outcomes across diverse referral problems. School consultation is often viewed as a preventive strategy for addressing a range of academic and behavioral problems that typically occur in the classroom (Begeny, Schulte, & Johnson, 2012; Bergan, 1977). Perhaps it is this view that has led many researchers to examine consultation’s impact on more than one type of problem within a single study (e.g., DuPaul et al., 2006; Fuchs, Fuchs, Bahr, Fernstrom, & Stecker, 1990; Jitendra et al., 2007; Ruble, Dalrymple, & McGrew, 2010; Sheridan et al., 2009). Although such a strategy increases studies’ external validity, it presents a measurement difficulty because client improvement must be assessed in relation to different target behaviors for each client. A similar issue arises when researchers attempt to cumulate consultation outcomes across studies in meta-analyses.

Multiple strategies have been employed in consultation studies to cumulate and characterize treatment outcomes across different referral problems. These strategies have included (a) “generic” rating scales, (b) “generic” observational measures, (c) goal attainment scaling (GAS), and (d) effect sizes calculated for individuals and averaged across cases.

Generic rating scales were used by Fuchs and colleagues in a series of studies concerning consultation as a prereferral strategy (Fuchs & Fuchs, 1989; Fuchs, Fuchs, & Bahr, 1990; Fuchs, Fuchs, Bahr, Fernstrom, et al., 1990). In the Fuchs studies, teachers and consultants together created behavioral descriptions of a child’s classroom difficulties and then the teacher rated each child’s target problem in terms of its manageability, severity, and tolerability on three 5-point scales. Scores on the three scales were summed to examine changes in consultees’ perceptions of whether target behaviors had changed from pre- to post-testing for the treatment or control groups. The summed ratings showed good internal consistency, but no other psychometric characteristics were reported. Noell et al. (2005) used a similar strategy based on the Fuchs’ measures.

More recently, generic rating scales for teachers and parents that are applicable across a range of referral problems and have more extensive information about
their psychometric characteristics have become available. For example, Fabiano et al. (2010) used the Impairment Rating Scale (Fabiano et al., 2006) as one outcome measure in their school consultation study. Similar to the measure used in the Fuchs study, teachers rated the severity of children’s classroom problems using one-item scales that assessed several key domains of school functioning (e.g., peers, academic progress). The scale has reliability and validity data and there is an established cut-off indicating impairment in a domain (Fabiano et al., 2006).

Weisz et al. (2011) have developed a measure called “Youth Top Problems” to track changes in problems identified as most important by children and parents during child treatment. Parents and children separately identify the problems that concern them the most and then rate their severity on a scale of 1 to 10. Both parents and children separately also rate their top three problems. During treatment, children and parents are asked to rate the severity of the top three problems they have identified on a weekly basis. Data on the measure’s reliability, validity, and sensitivity to change over time are quite encouraging, and slopes of change over time on the individually tailored items paralleled slopes from an established rating scale of internalizing and externalizing problems in youth (Weisz et al., 2011). A similar idiographic approach might be adapted for use in school consultation research.

The three Fuchs studies mentioned above (Fuchs & Fuchs, 1989; Fuchs, Fuchs, & Bahr, 1990; Fuchs, Fuchs, Bahr, Fernstrom et al., 1990) also used the third strategy, an observational measure that allowed results to be cumulated across target behaviors. For each student participating in the consultative treatment, the student and two randomly selected same-sex peers were observed before and after consultation. The mean percentage of intervals in which students exhibited the particular behavior that had been targeted for intervention in the treated student was then recorded, and the discrepancy between the target pupil’s behavior and his or her peer’s behavior was used as the measure of the target pupil’s functioning. Although the observers were not “blind” to the treatment condition of participants, independent coders who did not know participants’ treatment condition assignment were used in reliability checks and agreement between both types of observers was high (Fuchs, Fuchs, Bahr, Fernstrom et al., 1990).

The third strategy, goal attainment scaling (GAS; Kiresuk, Smith, & Cardillo, 1994), has been used in several consultation studies (e.g., Kratochwill et al., 1995; Ruble et al., 2010; Sheridan et al., 2009) to cumulate results across consultation cases that addressed different referral problems. With this technique, individual goals for each client are scaled on 5-point scales, typically ranging from −2 to +2. Negative scores (−2, −1) indicate poor outcomes and positive scores (+1, +2) indicate good outcomes. The midpoint of each scale (0) represents client functioning at baseline (no change) or the expected outcome at the close of treatment, depending on which variant of GAS is used (Hughes et al., 2001; Kiresuk et al., 1994). Persons completing the scales for clients can be independent judges or persons participating in the treatment under study. There is support for both the reliability and validity of GAS (see Hurn, Kneebone, & Cropsey, 2006; and Schlosser, 2004; for recent reviews of the methodology), although concerns have been raised about scale weighting procedures and whether the scale points should be treated as equal interval data (MacKay & Somerville, 1996), as well as the fact that low correlations...
between GAS and standardized measures of client outcomes are often observed (Schlosser, 2004).

The final strategy for cumulating outcomes across different types of target problems is the conversion of treatment outcomes to effect sizes. A variety of ways of calculating effect sizes are available that are appropriate for single-subject, within-subject group (i.e., repeated measures), and between group designs (Busk & Serlin, 1992; Durlak, 2009; Hunter & Schmidt, 2004; Shadish, Rindskopf, & Hedges, 2008; WWC, 2011). Multiple consultation studies have used effect sizes as the primary outcome measure when reporting on a series of single-subject case studies that addressed diverse referral problems, including Kratochwill et al. (1995), Sheridan, Eagle, Cowan, and Mickelson (2001), and Sheridan et al. (2009). Although effect sizes are commonly used in cumulating results across subjects in single case designs in many treatment areas, there is no expert consensus on which of the many ways of computing effect sizes is appropriate, and none of the present procedures yield effect sizes that permit valid comparisons of treatment effects across group and single-case designs (Shadish et al., 2008). WWC has indicated a preference for a regression-based estimator, or the comparisons of effect sizes calculated multiple ways to determine if the intervention effect appears similar across methods of determining effect size. WWC has also introduced a variant of effect size for group designs, the improvement index, which indicates the expected change in percentile rank for participants in the treatment group compared to the comparison group.

The four strategies just described differ in terms of their advantages and disadvantages for cumulating and characterizing outcomes across different referral problems. All offer the advantage of allowing outcomes to be cumulated within a study across diverse problems. In terms of rigor, not all of the strategies employ objective observers, but each could be modified to use objective observers rather than treatment participants.

Compared to the other strategies, the peer-discrepancy observational strategy (Fuchs & Fuchs, 1989; Fuchs, Fuchs, & Bahr, 1990; Fuchs, Fuchs, Bahr, Fernstrom et al., 1990) offers the advantage of providing information about the extent to which the consultation normalized client behavior compared to peers. However, the use of a discrepancy score introduces unreliability (Cone & Wilson, 1981). This problem may be lessened by the fact that the scores used in calculating the discrepancy are based on observations of the client and his or her peers. High correlations increase the measurement error and discrepancy scores that are based on two scores from different persons are less likely to be highly correlated than those based on two scores from the same person. Nevertheless, there is likely to be a correlation between the peer and target pupil’s behavior because they are observed in the same setting.

Effect sizes are widely used in characterizing the impact of treatment and cumulating results across subjects or studies (e.g., DuPaul & Eckert, 1997; Lipsey & Wilson, 1993; Weisz et al., 1995), and are now a standard way of characterizing study outcomes in research reports (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). However, caution is needed in using and interpreting effect sizes as a means of summarizing treatment outcomes across different referral problems. The magnitude of effect sizes are dependent on both the manner in which they are calculated and the type of study
design on which they are based (Durlak, 2009; Parker et al., 2005; Rosenthal, 2000). This fact has sometimes been ignored in reporting and interpreting the meaning of effect sizes within and across consultation studies (e.g., Reddy et al., 2000). Effect sizes that are based on repeated measures of the same person or group without the use of a control group will be inflated by autocorrelation relative to effect sizes calculated from between group designs (Busk & Marascuilo, 1992; Shadish et al., 2008). Thus, mean effect sizes from different designs must be reported separately (Durlak, Meerson, & Foster, 2003), and effect sizes calculated on single subject or within-group designs cannot be interpreted using Cohen's (1992) criteria for small, medium, and large effect sizes.

Summary

The limited breadth and quality of outcome measurement in consultation research has been a focus of criticism (e.g., Gresham & Noell, 1993). However, there are many examples of contemporary consultation studies that have sampled multiple domains and employed high quality measures. A multi-source, multi-method approach is recommended by two of the methodological sources. Although this approach may dilute the power of studies to detect treatment effects, current methods for adjusting for family-wise Type I error have reduced this risk over older procedures that applied overly stringent correction procedures.

FUTURE DIRECTIONS

From the first to second edition of this book and chapter, the impact of the evidenced-based treatment movement on the quality of consultation studies is clear. There are many examples in the current consultation literature of studies where extensive and high quality measurement strategies have been used across many of the facets of study design discussed here, including tracking all participants, assessing consultation procedural integrity and intervention plan integrity, and assessing and reporting client outcomes (e.g., DuPaul et al., 2006; Fabiano et al., 2006; Jitendra et al., 2007; Sheridan et al., 2012).

Weisz et al.’s (2005) characterization of the methodology used in youth psychotherapy research studies provides a useful yardstick for examining the state of current consultation research. Although direct comparisons are not possible because the consultation research methodology has not been reviewed as recently nor as comprehensively as the youth psychotherapy research, the similarity in methodological issues and weaknesses in the two treatment approaches is striking. Despite the recent progress in consultation research, it is clear that there are many aspects of research design and reporting where “best practice,” as it is defined by emerging cross-disciplinary standards for evaluating treatment studies, must be implemented more consistently by researchers investigating both psychological treatments. In addition, there are a number of areas where measurement technology is not yet well developed, particularly in the area of treatment integrity, and further research and development is needed.

Table 3.1 presents a number of short- and long-term suggestions for improving the quality of measurement in the consultation literature. The suggestions are
Table 3.1 Basic Considerations and Future Directions for Improving Measurement and Reporting in Consultation Treatment Studies

<table>
<thead>
<tr>
<th>Participant Characteristics</th>
<th>Future Directions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Participant Characteristics</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Basic Considerations</strong></td>
<td><strong>Future Directions</strong></td>
</tr>
<tr>
<td>Report the number of participants at each phase of a consultation treatment study and report the type of analysis used in reporting outcomes (e.g., intention to treat versus per protocol).</td>
<td>Develop more appropriate means of characterizing participants on diversity dimensions, such as acculturation and ethnic identity.</td>
</tr>
<tr>
<td>Report basic client demographic characteristics, including age, gender, ethnicity, socioeconomic status, and grade.</td>
<td>Move toward consensus in child clinical and school psychology on participant and setting characteristics that should be included in journal reports.</td>
</tr>
<tr>
<td>Report basic consultee and consultant characteristics, including age, gender, ethnicity, occupation, years of experience, and training in profession and consultation.</td>
<td></td>
</tr>
<tr>
<td>If results are intended to generalize to a clinical population, use established clinical criteria and formal measures to verify diagnoses.</td>
<td></td>
</tr>
<tr>
<td>When school-identified special education students are study participants, provide a statement of classification criteria used by the schools, and independently assess “caseness,” or describe the sample on key dimensions related to the disability classification.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consultation Process</th>
<th>Future Directions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Considerations</strong></td>
<td><strong>Future Directions</strong></td>
</tr>
<tr>
<td>Use a treatment manual to describe consultation procedures.</td>
<td>Refine the technology for assessing consultation procedural integrity through delineation of (a) unique; (b) essential, but not unique; (c) compatible, but neither necessary nor unique; and (d) proscribed consultant behaviors within each model of consultation.</td>
</tr>
<tr>
<td>Include measures of both consultation procedural integrity and intervention plan implementation that can be verified by independent observers.</td>
<td>Conduct comparative studies of the reliability, generalizability, and validity of consultation procedural and intervention plan integrity measures to provide better understanding of cost-effective strategies and sampling plans.</td>
</tr>
<tr>
<td>Continue developing consultant and consultee measures that assess competence as well as adherence. Design consultation procedural integrity measures that apply across models for differentiation of consultation models and assessment of elements common to all models.</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
organized around the three domains covered in this chapter. “Basic considerations” are actions that can be easily achieved in the short term given the current state of consultation treatment manuals and measures. “Future directions” are actions that will increase the rigor of consultation research in the long term, by improving our tools or increasing our understanding of measurement issues.

In conclusion, this chapter has covered a wide range of topics related to strengthening the research base for school consultation by improving the accuracy of measurement related to describing study participants, validating study process, and documenting treatment outcomes. Many topics related to measurement, such as the assessment of consultee skill development as a result of consultation, were not discussed. The length and complexity of this chapter, despite these omissions, is a testament to the difficulty involved in designing and conducting treatment research, particularly when studying an indirect treatment mode such as consultation.

Researchers face an unending battle to increase the signal-to-noise ratio in psychological treatment research (Hunter & Schmidt, 2004). If Wilson and Lipsey’s (2001) assertion is correct, and how a construct is measured contributes as much variance in outcome research as what is measured, then continued attention to measurement issues by consultation researchers is key to producing a corpus of high quality studies that document when and how consultation yields its promised benefits to youth at risk.

**NOTE**

1. In an “intention to treat” analysis, the last data point collected is carried forward in subsequent assessments. With baseline data comprising both the pre- and post-treatment assessments, the effect size would be 0.
REFERENCES


