UNIVERSITY OF CALIFORNIA
RIVERSIDE

Preliminary Validation of the Brief Measure of Intervention Quality Scale

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Education

by

Anacary Ramírez

September 2022

Dissertation Committee:
Dr. Wesley A. Sims, Chairperson
Dr. Asha K. Jitendra
Dr. Rondy Yu
Dr. William P. Erchul

The Dissertation of Anacary Ramírez is approved:

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

Preliminary Validation of the Brief Measure of Intervention Quality Scale

by

Anacary Ramírez

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, September 2022
Dr. Wesley A. Sims, Chairperson

An assumption in the school consultation literature is that an effective consultation process facilitates an effective intervention. However, although having an effective consultation process is a crucial component, it does not guarantee the development or selection of a high-quality, evidence-based, and effective intervention within an indirect service delivery model. The Brief Measure of Intervention Quality (BMIQ) was developed to address the shortage of assessments of consultation-derived intervention quality for use in consultation efficacy and effectiveness research. Specifically, the purpose of this study was to provide preliminary evidence for the validity argument of this measure. Using an arguments-based approach to guide development and preliminary validation work, it was proposed that the BMIQ would collect defensible data indicative of intervention quality. This study was conducted in three phases. These phases included the development activities, content validation activities, and a pilot study to assess the degree to which scores from the BMIQ provided validity evidence for its interpretations and uses. In the end, the BMIQ represents a

promising instrument for the evaluation of interventions developed within a consultative process.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE: INTRODUCTION

School psychologists engage in a variety of tasks including referring, testing, developing interventions, and assisting in the development of prevention programs (Bahr et al., 2017). However, although many school psychologists report spending most of their workday on traditional assessment activities, surveys consistently indicate that consultation is one of the most preferred and valued professional activities for school psychologists (Bahr et al., 2017; Sullivan & Long, 2010). Broadly, consultation is an indirect service delivery model that can be used to assist consultees to enact change (Erchul & Martens, 2010). It is considered "indirect" because the consultant typically does not engage in direct work with the client and instead provides support to the consultee to problem-solve an issue at hand. One goal of this process is for the consultant to develop the necessary skills to problem-solve similar situations in the future without assistance from the consultant. Although this process can differ depending on the consultative model used (e.g., behavioral consultation, conjoint behavioral consultation), in schools, the general process typically adheres to a general problem-solving framework that includes relationship building, problem identification, problem analysis, intervention implementation, and program evaluation (Erchul & Martens, 2010).

Problem-solving has been considered the essence of consultation because it helps drive the development of school-based interventions as well as how they are implemented and evaluated (Zins & Erchul, 2002). Accomplishing the problem-solving process involves a series of tasks that includes relationship development, identification and analysis of the problem, intervention development/selection, intervention

implementation, evaluation of the intervention effectiveness, and follow up (Erchul & Martens, 2010). Each task is embedded with specific goals to achieve that ultimately conclude with the evaluation of the client's progress or lack thereof. In research, it is also common to evaluate the social validity and acceptability of the consultative process. This is frequently done using post-consultation questionnaires to assess various aspects of the consultative process including the acceptability of the intervention and the consultant themselves.

school consultation is generally considered both an effective and efficacious model of service delivery both in vivo and via teleconferencing technology, but largely absent from this literature are mechanisms to account for intervention quality within consultation activities (Bice-Urbach & Kratochwill, 2016; Erchul & Sheridan, 2014; Medway, 1979; Sheridan et al., 1996). Intervention quality in studies is often assumed rather than explicitly measured to strengthen internal validity of study findings. This is to say that, within consultation research, investigators assume that the consultation process will yield a high-quality intervention for clients. However, this assumption cannot be guaranteed. Because the school consultation process involves multiple interdependent components, a breakdown in any of these aspects could result in negative or null outcomes. As such, assessment may offer a potential solution for improving the internal validity of research evaluating the effectiveness of school-based consultation and intervention quality.

The Brief Measure of Intervention Quality (BMIQ) was developed to address the shortage of assessments of consultation-derived intervention quality for use in

consultation efficacy and effectiveness research. Specifically, as with any new measure, this study will serve to provide preliminary evidence for the validation argument of this measure. Using an arguments-based validity approach, it was proposed that the BMIQ would efficiently collect defensible data indicative of intervention quality for use to strengthen a) the internal validity of efficacy and effectiveness evaluations (i.e., empirical research) of school-based consultative services or b) the vetting and adoption process of interventions in applied school settings. To begin the accumulation of such evidence, this study examined whether the BMIQ captures the desirable "qualities" of an intervention, refine the number and organization of the individual items that make up this measure, and whether scores are reliable across raters. This pilot study began the validation process for the BMIQ through the accumulation of preliminary validity evidence to address the scoring, generalization, and extrapolation inferences underlying the proposed interpretations and uses for the BMIQ and serves as rationale for continued empirical validation efforts.

CHAPTER TWO: LITERATURE REVIEW

The following literature review will describe several critical aspects that are pertinent to the process of school consultation. First, an overview of consultation will be provided, including its definition and primary models used in schools. Next will be a discussion of the school consultation problem-solving model, followed by its demonstrated efficacy or effectiveness within schools. Then, the literature review will highlight the importance of evaluating the quality of interventions generated within the consultative process. Finally, a definition of intervention quality and its relation to the BMIQ will be provided, which is the aim of the investigation.

Consultation in Schools

One of the primary roles of a school psychologist includes working with teachers, administrators, students, and parents/guardians to create a safe, healthy, and supportive learning environment (NASP, 2020). With time, the responsibilities of school psychologists have evolved from an emphasis on referring and testing to one that includes several different responsibilities including data-based decision making, development of intervention and prevention programs, assessment, and consultation (Bahr et al., 2017). This cultural shift that includes an increasing devotion to consultation services involves not just consulting with teachers, but the potential to consult with parents/guardians or at the system level with an emphasis on program development (e.g., implementation of Multitiered System of Support [MTSS] district wide). In fact, consultation is considered a fundamental role for school psychologists that "permeates all aspects of service delivery" and is expected to expand expeditiously as the field continues

to shift away from an emphasis on assessment-related activities (NASP, 2020; Sheridan & Gutkin, 2000). Although many school psychologists report spending most of their workday on traditional assessment activities, surveys consistently indicate that consultation is one of the most preferred and valued professional activities for school psychologists (Bahr et al., 2017; Sullivan & Long, 2010).

Consultation

Broadly, consultation is an indirect service model that helps assist a person, group or organization with a delineated problem and change efforts with the guidance of a specialist (Erchul & Sheridan, 2014). There are several aspects of consultations that separate it from other helping processes including teaching and supervision. These include: (a) its triadic nature; (b) coordinate and nonhierarchical relationship between the consultant and consultee; (c) direct focus on work-related problems; (d) responsibility for the client's welfare; (e) consultee's freedom to accept or reject guidance from the consultant; and (f) confidentiality (Erchul & Martens, 2010). Within schools, consultation is a process that includes a school psychologist and teacher working together to remediate a specific student problem. More specifically, Erchul and Martens (2010) defined school consultation as:

> a process for providing psychological and educational services in which a specialist (consultant) works cooperatively with a staff member (consultee) to improve the learning and adjustment of a student (client) or group of students. During face-to-face interactions, the consultant helps the consultee through systematic problem-solving, social influence, and professional support. In turn, the consultee helps the client(s) through selecting and implementing effective school-based interventions. In all cases, school consultation serves a remedial function and has the potential to serve a preventive function. (pp. 12-13)

That is, the school consultant, typically a school psychologist, works with a consultee (e.g., teacher) by providing guidance and expertise so that the consultee may deliver a cooperatively developed plan or intervention to the client (e.g., student). Different variations of consultation can be employed in schools, all of which conceptualize these aforementioned roles in slightly varied manners, but typically adhere to a general problem-solving framework that includes relationship building, problem identification, problem analysis, intervention implementation, and program evaluation. Models of consultation frequently used in schools include mental health consultation, behavioral consultation, conjoint behavioral consultation, and organization development consultation (Erchul & Fischer, 2018).

Mental Health Consultation

Mental health consultation (MHC) began as part of an early effort by psychiatrists, most notably Gerald Caplan, to move away from traditional psychotherapeutic approaches towards a model of prevention (Sandoval, 2014). MCH has a strong emphasis on intrapersonal/person-centered issues and unconscious motivations for behaviors, likely due to its psychoanalytic theory influences (Erchul & Martens, 2018). More specifically, MCH is a process of interaction between two professionals (a consultant and a consultee) where the consultee requests the assistance of the consultant in a work-related issue that is within an area of the consultant's expertise to improve the functioning of a client with whom the consultee works (Caplan & Caplan, 1993). Within this model, the consultative relationship is voluntary and nonhierarchical, and is initiated by the consultee and is intended to improve the consultee's functioning with a client, and

to increase the consultee's professional skills in handling similar problematic situations in the future.

There are four overlapping types of MCH: client-centered case consultation, consultee-centered case consultation, program-centered administrative consultation, and consultee-centered administrative consultation (Caplan, 1970). These types of MCH differ from each other in terms of whether their focus is on individual cases or programs, and its emphasis on prevention or remediation. For example, consultee-centered mental health consultation focuses on the characteristics and ideas of the consultee that are contributing to their challenges at work in order to lead to improvements with the client and other clients in the future. The goal is to develop "a new way of conceptualizing the work problem so that the repertoire of the consultee is expanded and the professional relation between the consultee and the client is restored or improved" (Lambert, 2004).

Behavioral Consultation

The Behavioral Consultation (BC) model, originally proposed by Bergan in 1977, is an indirect service delivery model that involves a cooperatively, voluntary, and confidential relationship between a consultant (e.g., school psychologist) and a consultee (e.g., teacher) to address a client's (e.g., student) problems and increase the ability of the consultee to problem-solve a similar issue in the future (Bergan, 1977; Martens et al., 2014). In school settings, although the school psychologist may have little to no contact with the student, BC is considered well suited for addressing school-based problems because of its client-centered, problem-solving focus (Erchul & Martens, 2010). Moreover, this delivery method can be much more cost effective than direct services

because it allows the school psychologist to potentially impact more children than she/he/they would when delivering direct services (Sheridan et al., 1996).

The school-based BC model combines strategies and tactics of behavior analysis with a problem-solving approach that is used to develop evidence-based intervention plans and treatment outcomes (Erchul & Martens, 2010). The effectiveness and acceptability of the intervention is taken into particular attention by this model and is a viable model for delivering psychoeducational services through a multi-tiered system of support (MTSS; Erchul & Ward, 2016). BC uses a four-stage problem-solving model to increase the probability of creating an effective intervention. The first stage in the process is known as Problem Identification. During this stage, the school psychologist works with a teacher to define the problem behavior to get an estimate of the frequency, duration or intensity of the problem behavior and to identify potential antecedents and consequences (Gresham, 1982). The second stage is known as Problem Analysis. This stage involves validating the existence of a problem, identifying factors that influence the problem solution, and developing a plan with the teacher to address the problem behavior. The third stage is Plan Implementation, which involves the implementation of the intervention that is designed or selected by the school psychologist and teacher in a cooperative manner. Finally, the last stage of school-based BC is Plan Evaluation. Plan evaluation consists of assessing the data collected during consultation and the intervention process to see if the intervention was successful. Next steps are also discussed.

Conjoint Behavioral Consultation

Conjoint behavioral consultation (CBC) is an expansion of school-based behavioral consultation. It is a strengths-based, problem-solving, decision-making model of service delivery in which parents, teachers, and other caregivers work collaboratively to promote positive and consistent outcomes related to the child or adolescents' academic, behavioral, and social-emotional development (Sheridan & Kratochwill, 2008). In CBC, a consultant works with both the teacher and the parents interdependently and simultaneously (i.e., conjointly). With the active involvement of a trained consultant, the purpose of CBC is to facilitate collaborative work among individuals who play a significant role in a child's life. CBC allows for a collection of data across settings, which may enhance generalization and maintenance of the treatment. CBC makes several assumptions within its model. First, it assumes an ecological behavioral perspective to problem solving (Sheridan, 1997). That is, the home-school relationship is viewed as an interactive and cooperative triadic relationship. Moreover, it assumes that collaborative problem solving will create the greatest benefits (Sheridan, 1997). Finally, it assumes that all persons involved will be willing to participate, get along, and share information with one another to produce helpful insights and considerations regarding the child.

The stages of CBC parallel the four stages of BC, but with an added parent component. The first stage of CBC is the Conjoint Problem Identification (CPI) stage (Sheridan et al., 2014). This stage is operationalized through the CPI interview and therein the consultant works with the parent(s) and teacher to identify the student's needs, operationally define the behavior, determine factors that contribute to the behavior in

both settings, and define treatment goals and progress monitoring procedures. The second stage, Conjoint Needs Analysis (CNA), consists of another interview conducted by the consultant with the parent(s) and teacher. This interview is used to evaluate the baseline data collected, which are used to determine variables that influence the behavior and to develop a meaningful plan to address the behavior across the settings. Immediately following the CNA stage is the Plan Implementation stage, which consists of parent(s) and teacher implementing and monitoring the intervention that was developed. The final stage is Conjoint Plan Evaluation (CPE). The CPE stage consists of an interview by the consultant with the teacher and parent(s) to evaluate the effectiveness of the intervention and to determine the future course of action (e.g., continuation, termination, planning for maintenance, and follow-up).

Organization Development Consultation

Organization development consultation (ODC) is intended to have an impact on large groups, such as an entire school or school system, rather than an individual to improve communication skills, negotiate goals, and to reduce and resolve conflict (Bergan & Kratochwill, 1990). This model of consultation allows school consultants to share their expertise and knowledge with a large number of individuals within a system and extend their impact (Erchul & Fischer, 2018). Within ODC, the role of the consultant is to facilitate development of skills and activities, rather than intervene directly within the system or organization, to introduce new principles and practices into the organization, with the goal of effecting self-change so that its members can function more effectively (Bergan & Kratochwill, 1990; Castillo & Curtis, 2014). Intervention

procedures and techniques typically include group sessions to improve communication skills, to negotiate goals, and to reduce and resolve conflict.

Efficacy and Effectiveness of Consultation within the Schools

Considering the importance of consultative services in schools, it is extremely important to document the efficacy and effectiveness of consultation within schools. *Efficacy* studies refer to the cause-and-effect relationships between an independent variable (e.g., consultation) and a dependent variable (e.g., behavior) and whether one variable has caused a change in the other (Rosqvist et al., 2011). The goal of efficacy research is to identify the conditions under which a treatment or intervention works, identify its components, and manualize the procedures in a way that it is easy to replicate (MacLeod et al., 2001). This requires special attention to experimental control or internal validity. *Effectiveness* studies, on the other hand, are based on "real world" applications and an emphasis on external validity; that is, these investigations provide information about how well a treatment is perceived to work in an applied setting (MacLeod et al., 2001). Within consultation, this requires an analysis of data collected under natural conditions without university-based support and is often done through anonymous retrospective surveys to evaluate whether the treatment works for those who select it (MacLeod et al., 2001). Fortunately, over the years, school consultation has received a substantial amount of empirical support through case studies, meta-analyses, methodological reviews, and randomized control trials for being both an effective and efficacious model of service delivery (Erchul & Sheridan, 2014; Medway, 1979; Sheridan et al., 1996).

Consultation Efficacy

Sheridan et al. (1996) conducted an extensive review of school consultation outcome studies published between 1985 and 1995. Her review included 46 studies, including articles in peer reviewed journals and dissertations. Twenty-one of the studies used a BC model, five studies used a MHC framework, and the remaining twenty articles used a different consultation model or did not specify one. Most of the consultation studies targeted behavioral concerns (33%), academic concerns (22%), and skill building concerns (15%). Primary measures used to assess these outcomes included rating scales and direct observation, and outcomes were coded as either positive, negative, or neutral. Results indicated that consultation was largely efficacious because it led to positive results in about 76% of the studies reviewed. Studies that implemented a BC model led to the most positive results, with 95% of the studies indicating at least one positive outcome. Of the five MCH studies reviewed, three reported at least one positive finding. These results were very similar to the finding of previous consultation outcome investigations (e.g., Kratochwill et al., 1995; Mannino & Shore, 1975; Medway, 1979).

Another example of the efficacy of school consultation includes Reddy and colleagues (2000). This meta-analytic review included 35 consultation studies with child and adolescent-level outcomes published between 1986 and 1997. BC produced large positive effects on clients and consultees, ODC yielded robust effects at the system level, and studies using a MHC model produced medium effects on consultees. School consultation was found to be particularly effective for clients with externalizing behavior problems and academic difficulties. Large effects were found for consultee's learning

new skills and techniques, a reduction of referrals for psychoeducational assessments, and increased use of psychological services. Similar positive effects have been found when the modality in which school consultation is delivered is modified. For example, Bice-Urbach and Kratochwill (2016) examined the impact of delivering school consultation services via videoconferencing (i.e., teleconsultation) that was designed to reduce disruptive behavior in students living in rural communities. Results indicated that disruptive student behavior improved, and teleconsultation was found acceptable by teachers. Research has indicated that CBC is also an effective and acceptable model of service delivery. For example, according to Sheridan et al. (2014), there were a total of 21 published studies that investigated the effects of CBC using experimental or case study designs evaluating various behavioral, social-emotional, and academic concerns. Since then, at least four other empirical studies have been published (i.e., Bellinger et al., 2016; Garbacz & McIntyre, 2015; Garbacz, Watkins, et al., 2017; Ohmstede & Yetter, 2015) that yielded similar findings to those reported by Sheridan et al., (2014).

Consultation Effectiveness

Studies evaluating the real-world application of school consultation include retrospectively evaluating the consultant's skills and the quality of the services provided are less prevalent in consultation scholarship. As just one example, MacLeod et al. 2001 investigated the effectiveness of school-based behavioral consultation. Their investigation involved 80 public school teachers who were asked to evaluate a recent consultation case using a consultant effectiveness scale, a measure of consultation quality, and an intervention outcome index. Results indicated that the consultant's skills

were rated as highly effective. Regarding consultation quality, 4 of 6 critical elements of problem solving were included in a large majority of the cases and two thirds of the cases saw improved student functioning.

Components of School Consultation

Factors Influencing Consultation Outcomes

As discussed, school consultation is an indirect service delivery model that can serve both a remedial function and preventive function for challenges consultees may encounter (Erchul & Martens, 2010). School consultation, drawing upon the strengths of both behavioral and mental health consultation, is driven by three interrelated tasks: social influence, problem solving, and support and development. *Social influence* is defined as "a change in beliefs, attitudes, or behaviors of a target influence, which results in the action or presence of an influencing agent" while *social power* is the potential for social influence to occur (Mintzberg, 1983 as cited in Erchul & Martens, 2010). The social influence task involves influencing the consultee's perceptions to promote changes in their behavior. Social influence strategies that are used depend largely on the consultant's understanding of the power bases and are used as needed during the con

Table 1

Raven's (1992, 1993) Differentiated Social Power Bases

Social Power Base	Definition
Positive Expert	Person A does what Person B says because Person B possesses knowledge or expertise in a specific area of interest to Person A.
Negative Expert	Person A does the opposite of what Person B says because they believe that Person B is thinking of their own best interests.
Positive Referent	Person A does what Person B wants because they want to be similar to or associated with Person B.
Negative Referent	Person A does the opposite of what Person B says because they do not want to be similar to or associated with Person B.
Impersonal Reward	Person A complies with Person B because they perceive that Person B is capable of delivering tangible rewards.
Personal Reward	Person A complies with Person B because they believe Person B will like or approve of him/her for complying.

Impersonal Coercion	Person A complies with Person B because they perceive that Person B is capable of delivering tangible punishments.
Personal Coercion	Person A complies with Person B because they believe that Person B will dislike or disapprove of him/her/them for noncompliance.
Direct Information	Person A complies with Person B because the information provided by Person A makes logical sense.
Indirect Information	Person A complies with Person B because they overhear from a third party that a certain course of action worked well in a similar situation.
Formal Legitimate/Position	Person A feels obligated to comply with Person B because Person B occupies a position of status or authority.
Legitimacy of Reciprocity	Person A feels obligated to comply with Person B because Person B has done something positive for them in the past.

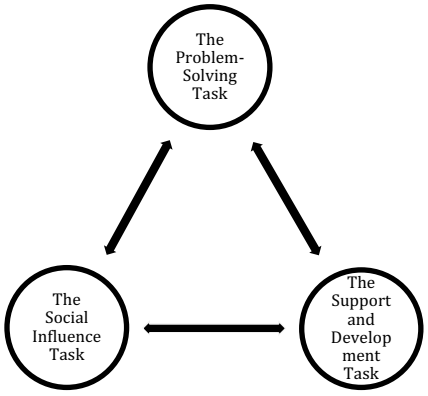
Legitimacy of Equity	Person A feels obligated to comply with Person B as a way of compensating for Person B's previous hard work.
Legitimacy of Dependence	Person A feels obligated to comply with Person B because Person B is unable to accomplish a certain action without his/her/their help.

Note. Adapted from Table 1 in Erchul, Raven, & Wilson (2004) and Table 3.1 from Erchul & Martens (2010).

Meanwhile, the problem-solving task involves a series of stages that includes relationship development, identification and analysis of the problem, intervention development/selection, intervention implementation, evaluation of the intervention effectiveness, and follow up (Erchul & Martens, 2010). The support and development task are to support the efforts of the consultee in dealing with crises that arise during their normal professional duties while facilitating the development of their skills (Erchul & Martens, 2010). These three tasks are linked because the problem-solving objectives of school consultation are accomplished through the consultant's utilization of a variety of interpersonal skills and techniques (i.e., social influence) while guiding and supporting the consultee through the consultative process (i.e., problem solving, support) to expand their repertoire of professional skills (i.e., development; see Figure 1; Erchul & Martens, 2010).

Figure 1

The Consultation Process



Note. Adapted from Erchul & Martens (2010).

Successful execution of these interrelated tasks should result in improved client outcomes and professional functioning of the consultee. For instance, if a teacher is apprehensive about using a Check-in Check-out intervention with a student due to their unfamiliarity with the intervention, a school psychologist, while using *direct informational social power*, may convince a teacher of the merits of using the intervention for behavior change. If the teacher continues to be apprehensive, the school psychologist may *problem solve* by providing brief part-time *support* while the teacher observes so they *develop* the skills to implement the intervention in the future and target the current client's behavioral goals. Additionally, these three tasks should be achieved while using a professional, consultative relationship. The core characteristics of this relationship includes: (a) a triadic alignment between the consultant, consultee, and client; (b) consultant-consultee relationship is characterized by cooperation and teamwork; (c) consultee's participation is voluntary; (d) the consultee has the right to reject the consultant's suggestions; (e) the consultee has an active involvement in

problem solving and plan implementation; (f) confidentiality of the information shared during the consultative interviews; (g) focus on professional, work-related issues; (h) pursuit of problem solving, social influence, and professional development goals; (i) emphasis on behavior analytic approaches; and (j) systematic evaluation of intervention outcomes (Erchul & Martens, 2010).

The Problem-Solving Task

All tasks of the consultation process are critical to achieve improved client outcomes and consultee skill development. However, the problem-solving process has been considered the essence of consultation because it helps drive the development of school-based interventions as well as how they are implemented and evaluated (Zins & Erchul, 2002). As discussed, accomplishing the problem-solving process involves a series of tasks which include relationship development, identification and analysis of the problem, intervention development/selection, intervention implementation, evaluation of the intervention effectiveness, and follow up (Erchul & Martens, 2010). The following sections will describe each phase in detail.

Relationship Development

The interpersonal relationship between the consultant and consultee plays a critical role in the effectiveness of consultation. Establishing a relationship is a precursor to the consultation process and begins with initial entry into the school or classroom and continues throughout the consultative process. This relationship requires trust, genuineness, and openness from both the consultant and consultee (Kratochwill et al., 2002). It also requires an integration of positive interpersonal skills and understanding to

maximize the consultant's effectiveness. Examples include acceptance through non-judgmental statements, openness, non-defensiveness, and flexibility. It also involves listening, talking, and nonverbal body language to establish and maintain constructive and professional interactions. During this stage, it is also critical that the consultant and consultee discuss and negotiate a mutually agreed-upon working contract or an oral or written understanding of what will happen during the consultative process as well as everyone's roles and responsibilities (Zins & Erchul, 2002). Consultants who have developed positive relationships with their consultees are likely to experience less resistance to the consultative process and intervention, find that their suggestions are readily accepted, increase the probability that the consultees will follow through with the intervention, and increase the effectiveness of the consultative process for consultees and clients (Kratochwill et al., 2002).

Problem Identification & Analysis

Problem identification is a key stage of consultation because it paves the way for the design and implementation of an effective plan. This stage involves identifying and defining the target problem in clear, concise, objective, and measurable terms, obtaining estimates of how often the behavior occurs, and initiating data collection (i.e., baseline; Erchul & Martens, 2010). Once the problem has been identified and defined cooperatively, goals can then be created. Goals are created during the problem analysis phase. This phase also involves using functional assessment strategies (e.g., functional behavior assessment, curriculum-based measurement, brief experimental analysis) as well as careful questioning and active listening to evaluate the behavior, academic, or social

competencies that need to be taught and the ecological context surrounding the problem. These two stages are key because ill-defined or misidentified problems may lead to an inappropriate or ill-matched intervention plan and wasted effort in attempt to solve the wrong problem.

Intervention Development & Implementation

An important goal of school consultation is to help consultees select/develop, implement, and evaluate intervention programs. This phase requires delineating the procedural details including who collects data, how the data will be collected, acquiring the required materials, and/or providing the required training to implement the intervention plan. The design and selection of a plan should be based on evidence-based interventions or practices. Resources that can be used for selecting an evidence-based intervention include reviews in scientific journals, What Works Clearinghouse, Intervention Central, and the Collaborative for Academic, Social, and Emotional Learning (Erchul & Fischer, 2018). In addition to empirical evidence, the design and implementation of an evidence-based intervention should also include attention to issues of acceptability, "buy in," effectiveness, client-match, and consultee skills and resources (Erchul & Fischer, 2018; Erchul & Martens, 2010; Kratochwill et al., 2002). An advantage within the intervention development phase is that it functions as a medium for individualizing evidence-based interventions to a specific client's behavioral, mental health, or academic needs and allows for the consideration of the consultee's perceived acceptability of the individualized intervention. Although this flexibility is advantageous, effectiveness of the interventions and factors that influence or hinder intervention

effectiveness (e.g., treatment integrity) has remained a top priority in research (Erchul & Young as cited in Erchul & Fischer, 2018; Erchul & Martens, 2010).

Treatment Integrity. Implementing an evidence-based intervention can take weeks or months. Within this time, the consultant and consultee typically engage in brief interactions so that the consultant can monitor intervention integrity, any possible side effects, and revise the plan, if needed. Keep in mind that the success of the evidence-based intervention selected or developed is largely dependent on the extent to which all the components of an intervention are implemented as intended (i.e., treatment integrity; Gresham, 1989; Sanetti & Kratochwill, 2009). Treatment integrity (or fidelity) is assessed using various dimensions including adherence, quality, and exposure (Dane & Schneider, 1998; Sanetti & Kratochwill, 2009). Within consultation, two levels of treatment integrity are important. These are procedural integrity and treatment plan implementation, which refer to the treatment integrity of the consultation process and the delivery of the intervention by the consultee, respectively (Noell & Gansle, 2014).

Monitoring treatment integrity within the consultative process is critical because it helps in determining whether the intervention was effective or responsible for any changes observed in the client. Methods for monitoring treatment integrity include self-report, direct observation, semi-structured interviews, and permanent products (e.g., student work products; Bergan & Kratochwill, 1990; Erchul & Fischer, 2018). For example, Bergan and Kratochwill (1990) provide a checklist for measuring procedural integrity when using a BC model which includes measuring the content and processes of a consultant's verbalizations. Treatment plan implementation forms can be part of the

intervention packet (see Sanetti et al., 2011) or developed by the consultant. Additionally, feedback during this phase has been found to considerably improve intervention integrity and the consistency of this effect is strengthened when data are presented to the consultee in graph formats (Noell & Gansle, 2014). Thus, ongoing support and feedback for the consultee's efforts is crucial.

Evaluation & Follow up

Plan evaluation and follow-up phases involve a series of intercorrelated tasks. These include establishing a sound basis for interpreting the outcomes of the intervention, evaluating the effectiveness of the plan and consultative process, and facilitating generalization of the new skills. First, the consultant and consultee must decide if the goals that were agreed upon at the start of the consultative process have been met. This is completed by reviewing the data collected to establish whether a change has occurred in the targeted behavior or skill. Then, post implementation planning to help ensure that maintenance and generalization of the client's new skill occurs. It is possible that generalization will occur naturally, but it can be facilitated with careful programming, such as introducing the intervention in a new environment. Finally, the consultant and consultee discuss whether to end or resume the consultative process, and if goals were not met, restart the consultative process. In research, it is also common to evaluate social validity or acceptability of the developed intervention, and consultative process and procedures.

Acceptability. Within consultation, measuring consultee perceptions of *acceptability* can include measuring the perceived acceptability of the consultation

process and procedures, the modality that is used to provide the consultative services (e.g., provision of service via teleconferencing technology), and the selected/developed intervention (i.e., treatment acceptability). Research has generally established that school consultation is an acceptable model of service delivery for both teachers and parents, even when delivered virtually (Fischer et al., 2017; Freer & Watson, 1999; Sheridan & Steck, 1995). With regards to treatment acceptability, this refers to judgments by consultees about whether a treatment is fair, reasonable, intrusive or appropriate for a given problem, and consistent with notions of what the treatment should be (Kazdin, 1980). Although acceptability should not be the sole criterion for which an intervention is selected, it is one of many important predictors of treatment integrity and effectiveness (Briesch et al. 2013; Erchul & Martens, 2010). Some research examples of social validity and treatment acceptability include: (a) *Behavior Intervention Rating Scale (BIRS*; Von Brock & Elliott, 1987) which evaluates the consultees' perception of the effectiveness of the behavioral intervention developed in the context of consultation; (b) the *Intervention Rating Profile-15 (IRP-15*; Martens et al., 1985) which assesses the consultees' perception of treatment acceptability; (c) the *Usage Rating Profile - Intervention (URP-I*; Chafouleas et al., 2009), which is an assessment instrument designed to evaluate the usage of an intervention broadly across intervention types to assess whether a consultee would adopt and subsequently utilize an intervention over time; and (d) the *Consultant Evaluation Form (CEF*; Erchul, 1987). The *CEF* asks specific questions about the consultees' perceptions of the helpfulness of the consultant, the benefits of consultation, and the overall satisfaction with the experience.

Consultation Derived Interventions

Numerous discrete tasks and goals are noted with consultative service delivery approaches, but broadly the goals of school-based consultation are twofold: a) address a challenge or issue facing a client through collaborative problem-solving strategies; and b) to increase a consultee's skills in handling similar problems in the future (i.e., multiplicative benefits of consultation; generalization). As noted previously, in the identification and analysis phases of consultation, the consultant and consultee must decide whether the conditions or skills should be considered for intervention and which conditions, or skills should be changed. From this point, consultation seeks to identify or develop mechanism(s) through which conditions or skills will likely improve. Often, these mechanisms take the form of changing the conditions surrounding the targeted outcome (e.g., achievement, behavior), helping or teaching the client skills needed to produce change, or both (Nadeem et al., 2013). Literature loudly and clearly advocate for engaging in interventions that are evidence-based, but also notes persistent barriers to use of such interventions in schools in favor of less efficacious, more easily accessible practices (Forman et al., 2009; Powers et al., 2010). In other instances, intervention efforts are more conducive to development rather than identification (e.g., behavior support plan versus Check in Check Out). Given barriers to adoption of evidence-based practices in schools as well as unique difficulties that prohibit use of a preexisting intervention, it is imperative to evaluate the quality of a selected or developed intervention. Unfortunately, few formal methods for evaluating intervention quality, across several underlying characteristics or factors, are noted in available research.

Intervention Quality

Research has established school-based consultation as an effective practice, but largely absent from this literature are mechanisms to account for intervention quality within consultation activities. This is to say monitoring procedural integrity, treatment integrity, and social validity as part of the consultation process is integral in consultation research and practice, empirical research in consultation does not appear to explicitly measure or account for the "quality" of the interventions developed during the consultation process. Intervention quality in studies is often assumed rather than explicitly measured to strengthen internal validity of study findings. This is to say that, within consultation research, investigators assume that the consultation process will yield a high-quality intervention for all participants (i.e., consultees and clients). When not assumed, consultation research has addressed this potentially confounding factor (i.e., intervention quality or inconsistence) by holding all interventions consistent across clients. One of the purported strengths of consultation, is the ability to tailor interventions to individual consultee and client needs (Erchul & Martens, 2010). By holding interventions constant through intervention predetermination, it could be reasonably argued that researcher is no longer evaluating consultation, but rather are evaluating the effectiveness or efficacy of the selected intervention.

A potential solution for the existing limitation, may lie in the use of assessment to establish and hold constant levels of intervention quality within consultation research. An assessment derived metric of intervention traits that increase its likelihood of success (e.g., evidence-base, acceptability) would allow the particulars of an intervention to vary

(e.g., target behavior, reinforcement schedule, reinforcers, length of direct instruction), while also ensuring an overall level of equity across interventions. Unfortunately, a review of available literature identified no existing assessment of intervention quality, or the intervention characteristics and implementation procedures (e.g., intervention match, cultural appropriateness, acceptability, resource requirements, evidence-base, clarity of the procedures). This is likely because of the assumption that an effective consultation process facilitates the development or identification of an evidence-based, effective intervention (Erchul & Martens, 2010). While intuitive and potentially more likely when interventions are developed (e.g., behavior intervention plans), adoption of a high-quality, effective intervention within consultation is not guaranteed and should not be assumed.

Although having an effective consultation process is a crucial and necessary component of developing or selecting an appropriate, evidence-based intervention within an indirect service delivery model, it may be insufficient in some instances. Using a school consultative process facilitates the problem-solving process; however, this process does not guarantee that there will be positive or observable changes on the targeted behavioral or academic goals, especially considering that the measured "effectiveness" or "ineffectiveness" of the school consultative process is moderated and mediated by various factors (e.g., relationship quality, acceptability, intervention integrity, intervention appropriateness; Frank & Kratochwill, 2014). Because the school consultation process involves multiple interdependent components, a breakdown in any of these aspects could result in negative or null outcomes. Thus, making it difficult to

differentiate a good consultative process with poor intervention implementation or intervention match versus a good consultative process with good intervention implementation and intervention match.

Intervention Quality Defined

Although there is a current lack of focus on intervention quality within consultation research, there are some clear indications as to what constitutes a "good" intervention. To start, a foundational characteristic of a good or quality intervention is whether it is based on evidence (Sanetti et al., 2014). Evidence-based interventions are strategies, curriculums or manualized programs that have been shown, in controlled research studies, to be efficacious in improving student outcomes (e.g., achievement or behavior; Sanetti et al., 2014). In other words, these interventions have been shown to work within their prescribed circumstances (e.g., settings, populations), when implemented as intended. Nevertheless, as previously discussed, it is not enough to select a good, evidence-based intervention, it is also critical to implement the steps of an intervention as they are intended to be implemented or, at the very least, the steps that have been identified as critical (Sanetti & Kratochwill, 2009). Research suggests that implementing an evidence-based intervention with fidelity is integral to achieve the desired behavioral or academic outcomes of an intervention. To do so, the critical steps and required resources (e.g., time, money, personnel) need to be carefully delineated (i.e., procedurally clear) in advance so that the intervention is useable and so that consultees can decide if it is economical for them to use (Sanetti et al. 2014; Johnson et al., 2018).

Whilst selecting an evidence-based intervention and implementing it as intended is vital, it is equally imperative that it be appropriate to the target population (e.g., addresses the target behavior, culturally sensitive; Domenech Rodríguez et al., 2011). As just one example, evidence-based interventions that have been culturally and sensitively adapted to diverse populations show much greater effects than traditional treatments (Domenech Rodríguez et al., 2011). This highlights the importance of quality interventions needing to be culturally responsive to the students with whom they are being implemented with. In addition to being culturally sensitive, adaptive, or appropriate, research suggests that good interventions must be readily accepted (e.g., fair, reasonable, appropriate for a given problem; Erchul & Martens, 2010; Kazdin, 1980). Although the level of acceptability does not impact the actual implementation of an intervention, it does predict treatment integrity and how effective an intervention is (Erchul & Martens, 2010). Additional qualifications discussed in the literature include whether the intervention is differentiated and targeted to the problem at hand (i.e., functional match; Gersten et al., 2008; Sanetti & Gritter, 2010). As just one example, Ingram et al. (2005) found that function-based behavior intervention plans were more effective at enacting behavior changes than plans that were not function based. These positive effects hold across diverse student populations and educational settings (Goh & Bambara, 2012).

Definition

Similarly, in the health sciences, broad characteristics that are used to determine "quality" programs or "quality" interventions include whether the intervention is feasible,

acceptable, based on theory, addresses the problem, and matches the target group (e.g., demographics, culture; Molleman et al., 2006). Additional aspects evaluated include the characteristics of the intervention itself, resources needed for implementation, and how clear the procedures of the intervention/program are. While keeping these aspects in mind, for the purpose of this study, a quality intervention was defined as one that is culturally sensitive/adaptive/appropriate to the target population, functionally matched to the problem at hand, procedurally clear to implement, based on evidence, economical for the implementer(s), and acceptable/feasible to implement. As such, a quality intervention that is culturally sensitive/adaptive/appropriate accounts for the socio-cultural values, beliefs, norms, status, and/or identity characteristics of the student/client. A quality intervention that is functionally matched to the problem is one that targets the underlying cause(s) of the problem (e.g., obtain peer attention, avoid written tasks due to skill deficit). Then, a quality intervention is one that is procedurally clear, which means it evaluates the precision, quality, and clarity with which all intervention components and procedures are delineated. A quality intervention is also one that is based on evidence meaning it has been shown to be efficacious in the literature. It is economical and describes and uses a reasonable amount of time and resources (e.g., money, materials). Finally, a quality intervention is one that is acceptable/feasible to implement suggesting that it evaluates the degree to which the procedures and components are feasible, reasonable, and manageable (i.e., implementers can implement the intervention as designed).

Brief Measure of Intervention Quality Development

School consultation is an effective indirect delivery service for improving client outcomes. One assumption within a consultation model is that it leads to the identification or development of an effective and appropriate intervention likely to remediate or improve the problem at hand. Based on this assumption, to evaluate consultation outcomes in research, the intervention is held constant across clients and settings. However, beyond holding all interventions consistent across clients, no mechanisms have been developed to assess intervention quality (e.g., intervention characteristics, traits, evidence base, and implementation procedures) to limit the potentially confounding impact of intervention variability within consultation research. Assessment may be a viable option for improving the internal validity of research evaluating the effectiveness of school-based consultation. For example, two interventions developed through a consultation process could vary across the goals, treatment mechanisms, assessment procedures, or other traits, but if assessed and determined to be of equitable quality, internal validity is preserved if not strengthened. Unfortunately, no such assessments were identified in literature to facilitate such activities. The Brief Measure of Intervention Quality (BMIQ) was developed using identified indicators of intervention quality, to address the shortage of assessments of consultation-derived intervention quality for use in consultation efficacy and effectiveness research. However, a central component of any assessment is the ability to use it with confidence. This confidence relies heavily on whether the measure is viewed as valid and reliable.

Validation Framework

A crucial issue of any new measure, such as the BMIQ, is that of validity. Validity is best conceptualized as a process that includes the development phase of an assessment and the argument phase (Cook et al., 2015). More specifically, assessment development begins by explicitly stating the proposed uses of generated scores, while also anticipating the empirical evidence needed to justify the use of scores and the empirical studies by which such validity evidence can be accumulated. This approach, proposed by Kane (2013), is labeled the arguments-based approach to assessment validation. As noted, in this approach, assessment developers begin by explicitly presenting the proposed interpretation and use argument (IUA) of the new assessment. The purpose of the IUA is to evaluate the key claims, assumptions, and inferences of a measure and evaluate what can be validly interpreted from these revelations (Cook et al., 2015; Kane, 2013). Validation then becomes a process through which evidence supporting the proposed interpretations and uses is accumulated. Validity evidence is accumulated across scoring, generalization, extrapolation, and implication/decisional inferences or a series of assumptions spanning from score generation to interpretations and implications (Cook et al., 2015; Kane, 2013). Scoring inferences refer to the process of moving from an observed performance to an observed score and should include the scoring rules, rubric, and scoring procedures (Kane, 2013). In the context of measurement development, researchers collect content validity and reliability evidence to determine whether the instrument is appropriate for assessing what it intends to measure and whether it allows scores to be applied consistently and accurately (Pua et al., 2021).

Generalization inferences refer to the degree to which the assessment protocol represents all the theoretically possible clinical events or scenarios (Kane, 2013; Tavares et al., 2018). Within the context of measurement development, researchers collect scoring evidence to show whether the sample observations represent all possible observations, provide reliable estimates of the construct being measured, and account for most of the variance observed (Pua et al., 2021). To support the extrapolation inference, hypothesized constructs sampled are expected to be representative of competence in the wider domain (Kane, 2013). In other words, researchers examine how observation scores are related to other measures of interest and are not influenced by systematic errors that undermine extrapolation of scores (Pua et al., 2021). Finally, the implication inference refers to the process of moving from scores to making decisions based on those scores (Kane, 2013; Tavares et al., 2018). Researchers collect evidence to show whether the instrument achieves its intended goals and results lead to positive impacts in the real world (Pua et al., 2021).

The IUA approach uses these inferences to connect assessment (or observation) to proposed decisions or uses. Each inference can and should be addressed through multiple methods of empirical testing, or the accumulation of validation evidence. Addressing these underlying inferences takes the form of familiar efforts to establish the psychometric defensibility of an assessment (e.g., concurrent validity, convergent/divergent validity, construct profile, reliability). To this end, BMIQ development began by clearly articulating the interpretations and uses argument (IUA) for generated scores. The IUA for the BMIQ proposed that the measure will efficiently

collect defensible (i.e., psychometrically sound) data indicative of intervention quality for use as a means to strengthen a) the internal validity of efficacy and effectiveness evaluations (i.e., empirical research) of school-based consultative services or b) the vetting and adoption process of interventions in applied school settings.

Purpose of Study

The purpose of this study is to begin accumulating validity evidence to support the inferences underlying the BMIQ and to support continued scholarly efforts. To begin the accumulation of such evidence, this study will first examine whether the BMIQ adequately measures the "quality" of an intervention, whether the individual items that make up the quality constructs are appropriate to maximize findings, and whether it does so consistently across raters. Because the quality constructs are hypothesized, this study will use a factor analysis approach to evaluate the claims being made. The specific research questions include: (a) What dimensions (i.e., factors) of intervention quality emerge from a preliminary pool of BMIQ? (b) What BMIQ item combinations most efficiently and appropriately capture intervention quality? (c) Is the initially identified factor structure confirmed in a secondary sample of BMIQ ratings? and (d) Do retained BMIQ items display acceptable levels of reliability within scale?

CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY

Participants

Participants included 47 students, faculty, and practitioners from School Psychology, Education, Special Education, and closely related fields. Study participants were recruited from a large urban metroplex in the Southwestern United States. Of those who participated and volunteered to share demographic information, most identified as women (N = 28; 60%) and Hispanic (N = 22; 47%). Most participants were undergraduates (N = 22; 47%) in their senior year (N = 11; 23%) in the field of education (N = 14; 30%). 12 participants declined to provide demographic information in-whole or part. See Table 2 for a complete summary of demographic information. To incentivize participation, participants were entered into a raffle to win one of 20, $25 Amazon gift cards regardless of study participation.

Table 2

Participant Demographic Information

Category	Subcategory	Sample	
		N	%
Gender	Female	28	60
	Male	8	17
	Neutral/Non-conforming	2	4
	Unknown	9	19
Ethnicity	Hispanic	22	47
	Non-Hispanic	16	34
	Unknown	9	19
Race	Hispanic or Latino	15	32
	African American/Black	4	9
	Asian	7	15
	American Indian/Alaska Native	1	2
	Middle Eastern/North African	1	2
	White	8	17
	Unknown	11	23
Degree Status	Undergraduate	22	47
	Graduate	11	23
	PhD/PsyD/EdD	5	11
	Unknown	9	19

Years in School	Doctoral Year 1	2	4
	Doctoral Year 2+	6	13
	Masters Year 1	4	9
	Senior	11	23
	Junior	9	19
	Sophomore	1	2
	Freshman	1	2
	Unknown	12	26
Field of Study	Education	14	30
	School Psychology	10	21
	Psychology	9	19
	Music Performance	1	2
	Psychology/Education	1	2
	Sociology	2	4
	Unknown	10	21

Instruments

The Brief Measure of Intervention Quality (BMIQ) was designed to measure six hypothesized constructs research suggests contribute to intervention quality. Preliminarily hypothesized constructs included: a) Culturally Sensitive/Adaptive/Appropriate b) Functionally Matched, c) Procedurally Clear, d) Evidence-Based, e) Economical, and f) Acceptable. See Table 3 for a description of each construct. Initially, 39 BMIQ items were developed across the six hypothesized constructs. All items are scored using a unidimensional Likert-style rating system ranging from 0 to 5. Qualitative item anchors ranged from *Strongly Disagree* (0) to *Strongly Agree* (5). The Culturally Sensitive/Adaptive/Appropriate construct included ten items, the Functionally Matched construct included seven items, Procedurally Clear originally included six items, the Evidence-Based construct included four items, Economical included six items, and Acceptable included six items. Following a content validation activity, 16 items were removed (see description below). The resulting BMIQ included four items for the hypothesized Culturally Sensitive/Adaptive/Appropriate construct, five items for the Functionally Matched construct, three items for Procedurally Clear, four items for Evidence-Based, three items for Economical, and four items for Acceptable. A total of 23 items were used in the initial examination of the instrument's underlying factor structure. The preliminary BMIQ included 22 positively anchored items (i.e., higher scores are more desirable) and one negatively anchored item (i.e., lower scores are more favorable). The latter, Item 16, required reverse coding when scoring participant responses. Additionally, reverse coding this negatively anchored items facilitates

computation of a Total BMIQ Score, a score designed to be a metric for overall intervention quality.

Table 3

BMIQ Category Definitions

Subscale/Construct	Definitions
Culturally Sensitive / Adaptive / Appropriate	Items evaluate the degree to which the intervention accounts for the socio-cultural values, beliefs, norms, and preferred identity characteristics of the student/client (if/when applicable).
Functionally Matched	Items evaluate the alignment of the intervention with the underlying cause of the problem (e.g., obtain peer attention, avoid written tasks due to skill deficit).
Procedurally Clear	Items evaluate the precision, quality, and clarity with which all intervention components and procedures are communicated.
Evidence-Based	Items evaluate the connection of the intervention to theoretical and/or empirical evidence supporting its use.
Economical	Items evaluate the amount of time and resources (e.g., money, materials) necessary for implementation.
Acceptable	Items evaluate the degree to which intervention procedures and components are feasible, reasonable, and manageable (i.e., implementers can implement the intervention as designed).

Procedure

Study activities were completed in three phases. Phase 1 consisted of general study conceptualization, literature review, and preliminary BMIQ item development. A second phase included a formal content validation activity. The third phase encompassed primary study data collection activities. For this phase, a preliminary version of the BMIQ was completed by study participants. Data from this pilot administration was used to address primary research questions guiding this study and to collect initial validation evidence. Prior to data collection, an application for Human Subjects Research was completed and approved by the University of California, Riverside Institutional Review Board.

Initial BMIQ Development

The development phase, which also serves to address the scoring inference, followed a three-step process outlined by Lynn (1986). This process serves as a guide to the preliminary development and formatting of a new assessments across: a) domain identification; b) item generation; and c) instrument formation. Domain identification describes the dimensions or subdimensions that are identified through a comprehensive literature review. Item generation refers to the item development process for the dimensions or subdimensions identified. Instrumentation formation is used to denote the process in which the items generated are assembled in a useable format. In keeping with these steps, a comprehensive literature review was conducted to identify salient components of high-quality school-based interventions, or hypothesized BMIQ domains. This review revealed numerous discrete elements or characteristics empirical literature

has established as critical for intervention effectiveness. As noted earlier, some prominent characteristics linked to more effective interventions included evidence-base, acceptability, and functional match. Additionally, scholarly works identified several other aspects of interventions theorized to impact effectiveness. Literature hypothesizes the importance of culturally adaptive and appropriate interventions. Although intuitive, researchers have yet to empirically evaluate these theorized connections. Following this extensive literature review and identification of discrete characteristics of high-quality school-based interventions, assessment developers identified commonality or themes within identified characteristics. This resulted in six domains (noted above) BMIQ developers believed encompassed all identified characteristics of high-quality, effective school-based interventions: a) Culturally Sensitive/Adaptive/Appropriate b) Functionally Matched, c) Procedurally Clear, d) Evidence-Based, e) Economical, and f) Acceptable.

Somewhat inconsistent with the sequence presented by Lynn (1986), but more intuitive and efficient, BMIQ developers selected an instrument format, broadly, that aligns with the proposed BMIQ interpretations and uses. Consistent with the proposed interpretations, BMIQ developers sought to format BMIQ items in a manner that would be low inference, low cognitive load, and efficient while generating quantifiable scores. Cognitive load can be defined as a multidimensional construct describing the taxing of cognitive energies and systems (e.g., memory, attention, inferencing, semantic knowledge) that performing a particular task imposes on a person (Paas et al., 2003). To this end, developers approached item development anticipating the resulting instrument formation would adopt a Likert scale format. The advantages of Likert scales are well-

documented and easily adapted to lower inference, lower cognitive, and expeditious assessment or observation activities (Wu & Leung, 2017). Advantages included: a) data can be gathered rather quickly from large numbers of participants; (b) the data can provide highly reliable person ability estimates, (c) the validity of the interpretations made can be established through a variety of means; and (d) the data provided can be compared, contrasted, and combined with qualitative data-gathering techniques, (e.g., observations, open-ended questions, interviews).

With an anticipated Likert scale format, BMIQ developers constructed items for aligning with content or examples consistent with each of the domains identified during the literature review (e.g., Is the intervention is appropriately sensitive/adaptive/responsive to the student/client's racial/ethnic background? Is the intervention based on evidence?). Once generated, developers reviewed, discussed, and refined developed items. Additional efforts were made to improve the wording of the extant items and to generate additional items for each hypothesized factor. Items deemed redundant or repetitive were eliminated. Items were eliminated judiciously given planned, empirically driven item pruning (i.e., formal content validation activities, item reduction during factor analytic procedures). Ultimately, using the opinions and edits of experts (three school psychology faculty and two school psychology graduate students), 39 items were included in the formal content validation activity.

Lastly, generated items were used for formal instrument formation. A total of 39 preliminary items hypothesized to measure characteristics encompassed by the six domains identified in intervention quality literature were formatted into a rating scale.

Developers adopted 6-point Likert-style rating system ranging from 0 (*Strongly Disagree*) to 5 (*Strongly Agree*). A 6-point scale was selected to limit raters' ability to respond in a neutral and potentially less informative manner during future use of the BMIQ. Specifically, a body evidence suggests that some individuals may display a tendency to satisfice, or avoid the cognitive effort required to provide a genuine or satisfactory answer when responding to some assessment activities (e.g., evaluation ratings; (Johns, 2005; Krosnick, 2002; Nowlis et al., 2002). Additionally, as noted previously, a Likert-style format was selected given respondent's likely familiarity. As part of the instrument formation process, item wording was adjusted to align with the selected qualitative rating anchors (e.g., *Agree*, *Strongly Agree*, *Disagree*). Although generated using hypothesized domains as guides, items were not formatted to group together under these domains for the subsequent content validation activity. Such grouping could have unduly influenced participant rankings of items relative to hypothesized domains.

Content Validation

Given recommendations from McKenzie et al. (1999), five individuals served as the content validation panel. Panelists include doctoral students in school psychology (2), school psychology faculty (2), and a practicing school psychologist from a large research-intensive university and large suburban school district. Panelists were selected based on their background and expertise in consultation, assessment, and school-based interventions. The five panelists agreed to participate and were emailed a link to the content validation activity. Preliminary items were input into Qualtrics, a web-based

survey program. The Qualtrics survey asked participants to judge and sort or rank items across three considerations: a) primary domain, b) secondary domain, and c) importance to evaluating intervention quality. First, each judge was asked to conduct a subjective evaluation of each item relative to proposed domains and their definitions. The panelist then selected the domain to which they perceived the item to best align. Panelists were then tasked with selecting the domain for which the perceived the item to be the next best fit. Finally, judges provided an indication of the perceived importance/relevance of each item to evaluating intervention quality. Perceived importance/relevance was rated on a 5-point Likert scale. Rankings ranged from *Not at all important* to *Absolutely Critical*. See Appendix A for content validation activity example items.

The goals of the content validation were two-fold. First, results were used to inform preliminary item-domain assignment (i.e., content validity). This assignment served to conclude preliminary instrument formation. This is to say, panelist responses facilitated assignment of items to one of the six hypothesized domains. Items were then organized or formatted by domain for presentation to participants in the pilot administration. Second, panelist responses facilitated pruning of likely unnecessary items, those that displayed ambiguous domain alignment or were perceived as irrelevant to evaluating intervention quality. Based on panelist responses, a total of 23 items were retained subsequent to the content validation process. A priori rules were used to determine which items would be retained or deleted; that is, items were removed if (a) panelists agreement for item-domain alignment was not evident or easily resolved or b) the importance/relevance rating for an item fell below "important" across panelist reports.

Agreement across item-domain alignment across panelists was defined as all panelists selecting the same domain for either primary or secondary item alignment. For example, agreement was noted if all panelist indicated the same domain as displaying the best alignment with a given item. Additionally, panelist agreement would be noted if four panelists selected the domain "Functionally Matched" as the domain to which Item 7 aligned and the fifth panelist selected "Functionally Matched" the domain with the second-best alignment. In these instances, Item 7 would then be grouped with other items which panelists indicated alignment with the "Functionally Matched" domain for the pilot administration. If judges disagreed about the construct that the item measured but believed that item was highly relevant, items were kept, and an expert panel resolved the disagreement about the assignation to a construct via discussion. Ultimately, no items were removed due to nonagreement for domain alignment.

A total of 16 items were removed based on panelist perceptions of irrelevance to evaluating intervention quality. A cursory review of the items that were deleted from the BMIQ revealed that the items were worded similarly to items retained, suggesting items may have been deemed irrelevant based on panelist belief that targeted information would be obtained via other items. After item removal, the remaining 23 items were input into a Qualtrics survey, in which they were grouped by the domain indicated in the content validation activity.

Pilot Administration

To evaluate the factor structure underlying the BMIQ and to collect additional scoring, generalization, and extrapolation evidence, the measure was used by a group of

students and professionals in the field of school psychology, psychology, and education to evaluate example behavior interventions. Participants evaluated four separate behavior intervention plans (BIP) using the 23 items retained following the content validation activity.

Example BIP Development

To facilitate BMIQ completion, a total of 21 example BIPs were developed. BMIQ developers utilized materials from a variety of sources to develop case study-style BIPs for use in the pilot BMIQ administration. BIPs consisted of research generated, case study intervention plans that varied in length and quality to allow for variability in participant appraisal and rating. The principal investigator solicited example or case study BIPs from practitioners, trainers, and graduate students in school psychology and applied behavior analysis (ABA) programs. Once accumulated, BIPs were reviewed and edited to remove any potentially identifying information and any typographical errors. Additionally, a standardized demographic cover page was developed to accompany each BIP to provide participants additional contextual information to facilitate BMIQ responses. Demographic information for each of the fictional students included educational information (e.g., special education eligibility), parent information (e.g., level of parental involvement), target behaviors, hypothesized function(s), replacement behavior(s), and strategies. BIPs were converted to PDF format and quality checks for image clarity were conducted before pilot administration activities began. See Appendix B for an example BIP.

Pilot BMIQ Administration

The 23 items retained following initial content validation activities were paired with example behavior intervention plans (BIPs) in a web-based Qualtrics survey. This pairing resulted in 21 possible BMIQ-BIP combinations, or block of survey items. Of these 21 possible combinations, participants were randomly assigned (via Qualtrics) to complete four BIP evaluations. Qualtrics only presented participants with the four BIPs they were tasked with evaluating. Each randomly presented item block (i.e., BIP-BMIQ combination) in Qualtrics included a brief greeting and introduction to survey activities, more specific directions, followed by an example BIP, and the BMIQ (see Appendix C). Specific directions included a request to spend a minimum of five minutes reviewing each BIP before completing the BMIQ. Following approval from the Human Subject Internal Review Board, the BMIQ study was distributed via email using an information sheet that contained an anonymous Qualtrics link to the study information, consent form, and brief demographic questionnaire. Once a participant had provided consent, an individualized link to the study was emailed to the participant via Qualtrics. The information sheet provided a description of what the study entailed, and the time commitment involved. Individualized links allowed participants to complete ratings over multiple occasions. Generally, participants completed the study in one, approximately 30-to-60-minute session.

Data Analysis

Approach to Data Analysis

Generally, guiding research questions were addressed through factor analytic techniques. In applied research, factor analysis is one of the most commonly employed methods in psychometric evaluations of testing instruments (e.g., questionnaires; Floyd & Widaman, 1995). In the early stages of scale development, factor analysis allows researchers to examine the plausible number of factors (i.e., latent, unobservable variables/constructs) within a set of observed measures (i.e., items or indicators), determine if those items are reasonable indicators of the underlying construct being measured, and eliminate unnecessary items (Brown, 2006). Additionally, factors help account for the variation and covariation among a set of indicators (i.e., items) giving researchers a more parsimonious understanding of the construct being measured (Brown, 2006). Common factor analysis, which emanates from the common factor model (Thurstone, 1947), suggests that each indicator is a linear function of one (or more) common factors and one unique factor; therefore, the variance of each indicator is partitioned into (a) common variance (i.e., variance accounted for by the factor; communality or h^2) and (b) unique variance (i.e., combination of reliable variance specific to that indicator; u^2) and random error variance (i.e., measurement error; e; Brown, 2006). The main analyses based on the common factor model are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Although both analyses aim to investigate observed relationships among a group of items with a smaller set of latent variables, they differ fundamentally (Brown, 2006). EFA is data driven and does

not require a priori specification of the number of factors. Without specific instructions to do otherwise, an EFA computer procedure could theoretically generate all possible solutions (Kline, 2016). In CFA, the number of factors and the pattern of indicator-factor loadings are determined a priori. Thus, unlike EFA, CFA requires a strong empirical foundation to guide the specification and evaluation of the factor model. In the present study, although it was hypothesized that the BMIQ would contain six constructs, no previous empirical evidence was found to support or validate each quality indicator. As such, both EFA and CFA are proposed in the current study to explore the factors that emerge.

Exploratory Factor Analysis

Exploratory factor analysis was used to address the first and second research questions, which sought to identify the dimensions (i.e., factors) of intervention quality the underly the BMIQ and to identify item combinations that most efficiently and appropriately capture intervention quality. Prior to doing EFA, several assumptions must be met which include population size, evaluation of missing data, violations of normality, and appropriateness of the data (Watkins, 2018). Regarding population size, a basic assumption is that the sample is large enough for the analysis (e.g., sample size >100). Given the preliminary nature of the current study, this assumption is being violated. Regarding missing data, listwise and pairwise deletion methods tend to be the default methods in most statistical packages, although not recommended for large proportions of missing data (Baraldi & Enders, 2010). Given the sample size and small proportion of missing data, the data was deleted listwise before the start of the current analysis.

To address severe violations of normality, skew and kurtosis of the items were evaluated. Skew refers to the symmetry of the score distribution, whereas kurtosis is a measure of the height of the score distribution in relation to its width (Watkins, 2018). To reduce the possibility of skew affecting EFA results, all variables should be scored in the same direction. Item 16 of the BMIQ was reversed coded to meet this criterion. Furthermore, univariate skew and kurtosis can be problematic when they are excessively elevated and may not be appropriate for use with Pearson correlation (Curran et al., 1996; Watkins, 2018). In situations where skew and kurtosis are elevated, more robust correlation methods (e.g., polychoric, tetrachoric) are recommended and were considered in this study. Thus, in addition to reverse coding item 16, kurtosis and skewness were evaluated to determine any significant violations to normality (≥3.0; Brown, 2006).

Finally, to address the appropriateness of the data for factor analysis, both subjective and objective methods are available. One subjective method available is to examine the correlation matrix (Watkins, 2018). Generally, items with very low correlations (e.g., ≤ 0.30) and high correlations (e.g., ≥0.85) are considered less than desirable (Fabrigar et al., 1999; Hair et al., 2010; Watkins, 2018). Thus, to avoid including items with a low degree of shared variance and undesirable correlations, low inter-correlations across at least three items, and high correlations across at least three items were deleted. On the other hand, objective methods available include the Kaiser-Meyer-Olkin statistic (KMO; Kaiser, 1974) and Bartlett's test of sphericity (Bartlett, 1950; Watkins, 2018). Bartlett's test of sphericity statistically tests the hypothesis that the correlation matrix contains one on the diagonal and zeros on the off-diagonal suggesting

that it was generated by random data (Watkins, 2018). To justify the use of EFA, Bartlett's test should produce a statistically significant chi-square value. KMO is the ratio of correlations and partial correlations that reflect the extent to which the correlations are a function of the variance shared across all variables (Watkins, 2018). KMO values range from 0.00 to 1.00 and can be computed for the whole correlation matrix as well as each measured variable. KMO values greater or equal to 0.70 are desired and values below 0.50 are generally considered unacceptable, indicating that the matrix is not factorable (Watkins, 2018). To ensure factorability, KMO values were required to be above 0.50 overall as well as for each measured variable. Measured variables that did not meet the cut-off were eliminated. Bartlett's test of sphericity was required to be significant.

Estimating the Factor Model

After analyzing the data and its factorability, there are four procedural aspects of EFA that must occur and be reported. These procedural aspects include: (a) selecting a specific method to estimate the factor model; (b) selecting the appropriate number of factors; (c) selecting a rotation technique to foster interpretability of the solution; and (d) selecting a method to compute factor scores, if desired (Brown, 2006). First, there are several methods available for extracting the common factor model. Some examples include maximum likelihood (ML), principal factors, weighted least squares, minimum residual analysis, to name a few (Brown, 2006). The two most commonly employed methods are ML and iterated principal axis (PA; also known as principal factors; Watkins, 2018). ML is sensitive to multivariate normality and usually requires large sample sizes, whereas PA is a least-squares estimation method that makes no

distributional assumptions. For the current study, an iterated principal axis (PA) extraction method was employed because of its relative tolerance of non-normality, tolerance of small samples (<300), and demonstrated ability to recover weak factors (Briggs & MacCallum, 2003; Watkins, 2018).

Number of Factors to Retain

A crucial step during an EFA is deciding how many factors to retain. It has been proposed that under-factoring (i.e., retaining too few factors) is more of a severe problem than over-factoring (i.e., retaining too many factors), but both can severely compromise the validity of the model and its estimates (Brown, 2006). Kaiser's criterion (retaining factors with eigenvalues over 1.00) is commonly used for deciding the number of factors to retain; however, numerous studies have indicated that it should not be used because of its tendency to overestimate the number of factors (Fabrigar et al., 1999; Fabrigar & Wegener, 2012; Watkins, 2018). Alternate recommended methods include parallel analysis (Horn, 1965), minimum average partials (MAP; Velicer, 1976), and the visual scree test (Cattell, 1966). Parallel analysis uses eigenvalues obtained from the sample data against eigenvalues simulated from a data set of random numbers (Brown, 2006; Watkins, 2018). The visual scree test uses visual inspection to find the "elbow" or distinct break in the slope of the scree plot. Meanwhile, to compute MAP, a matrix of partial correlations is calculated after each principal component is extracted, and the average of the squared partial off-diagonal correlations is calculated for each of these matrices until common variance has been removed and only unique variance remains (Watkins, 2018). Some empirical research suggests that MAP and parallel analysis are the most accurate

empirical estimates of the number of factors to retain while scree is a useful subjective method to the empirical estimates (Velicer et al., 2000; Velicer & Fava, 1998). Unfortunately, no method has been found to be correct in all situations; thus, as recommended by Fabrigar & Wegener (2012) several methods were employed. As such, parallel analysis, MAP, and the visual scree test were used to determine the appropriate number of factors to retain. Factor interpretability and researcher judgment based on theory were also considered.

Rotation of Factors

Once the appropriate number of factors has been determined, it is important to rotate the factors to create a simpler and more meaningful solution (Brown, 2006; Watkins, 2018). There are two types of rotations: orthogonal and oblique. In orthogonal rotation, the factors are constrained to be uncorrelated while in oblique rotation that factors are allowed to intercorrelate (Brown, 2006). Brown (2006) indicated that orthogonal rotation (e.g., varimax) tends to be used most frequently perhaps because it is the default in most major statistical programs (e.g., SPSS) and because of its perceived interpretability. In oblique rotation solutions (e.g., promax, oblimin), factor loadings usually do not reflect simple correlations between the indicators and the factors unless there is no overlap. Moreover, because factors are allowed to intercorrelate, the correlations between indicators may be inflated. Yet, using an orthogonal rotation when correlations are expected may yield an inaccurate representation of the relationships in the model (Brown, 2006). Given that EFA is being used as a precursor to CFA, oblique

rotations are generally recommended because they are more likely to generalize (Brown, 2006).

Due to the nature of the constructs and research involving humans, it was assumed that factors would be correlated in this study; therefore, to provide a more realistic representation of the interrelated nature of the factors, a non-orthogonal or oblique rotation was employed (Fabrigar et al., 1999; Watkins, 2018). If the factors produced using an oblique rotation are truly not correlated, it will virtually produce the same solution as an orthogonal rotation (Brown, 2006). One of the most popular oblique rotations is oblimin and was used for this analysis (Jenrich & Sampson, 1966).

Factor Scores

Once the appropriate factor rotation has been determined, it is important to evaluate the factor loadings and factor correlations. Evaluating the factor scores are important because estimation of factor correlations provides information about the relationship between factors, it can highlight the existence of redundant items and factors. For example, factor intercorrelations above 0.90 imply poor discriminate validity and suggest that a more parsimonious solution should be considered. If the correlations between all factors are moderate, then a single higher-order factor may account for those relationships (Brown, 2006). As part of the factor score review, it is critical to consider the meaningfulness of the factors (i.e., conceptual/empirical relevance), whether there are poorly defined factors (e.g., factors with insufficient indicators), whether there are poorly behaved items (e.g., cross-loading items [an item has ≥ 2 significant loadings above

0.30]), the significance of the items (e.g., item loadings >0.30; communality >0.40), and internal consistency reliability (e.g., reliability >0.70) of the model.

Summary

In sum, procedural steps for EFA include: (a) examining data fit; (b) factor extraction (i.e., iterated principal axis); (c) factor selection (i.e., scree plot, parallel analysis, MAP); (d) factor rotation (i.e., oblique rotation [oblimin]); and (e) interpretation of factors (e.g., quality, meaningfulness). Given this, criteria for determining EFA adequacy are: (a) factors must have a minimum of three indicators; (b) skewness and kurtosis must be equal to or less than 3.0; (c) there must be no items with low inter-correlations (0.30) or high inter-correlations (0.85) across three items; (d) KMO values must be above 0.50 and Bartlett's test of sphericity must be considered significant; (e) item communalities must be 0.40 or greater; (f) indicators must have factor loadings greater than 0.30; (g) indicators must not cross-load; (h) internal consistency reliability must be at least 0.70; and (i) model must be theoretically meaningful.

Confirmatory Factor Analysis

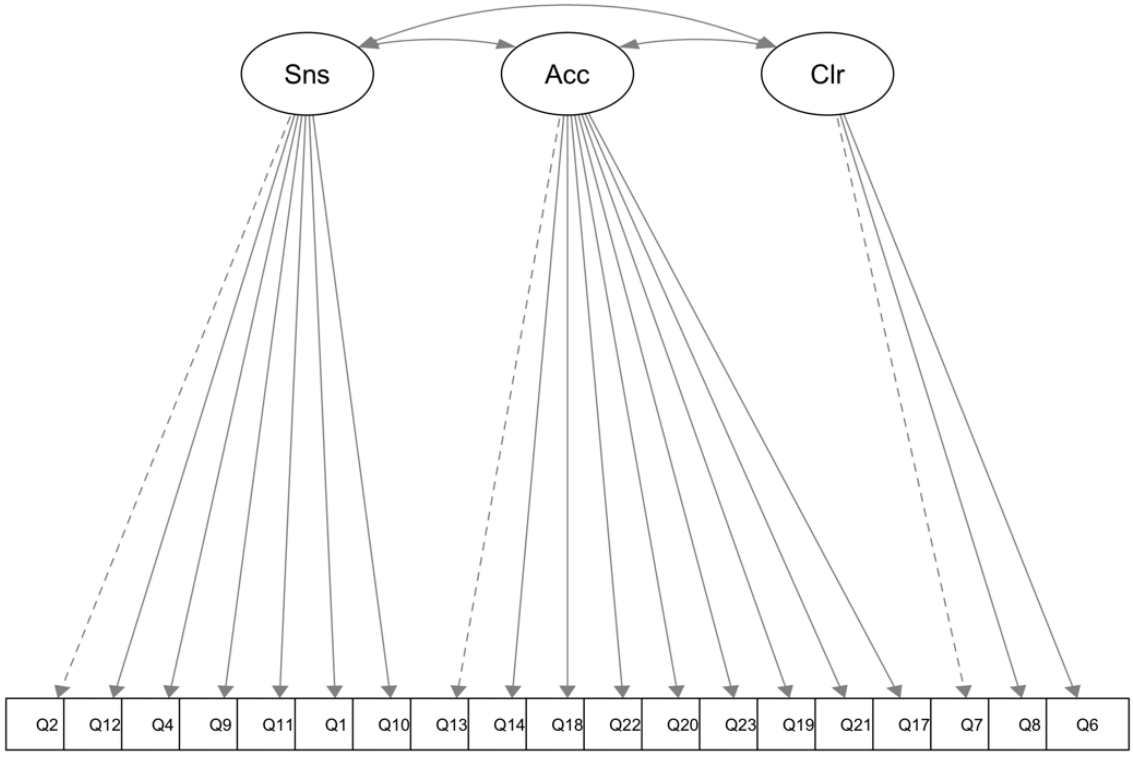
To address the third research question, which sought to confirm the intervention quality constructs identified by the EFA, CFA was used. Much like EFA, CFAs are used to identify latent factors that explain the variation and covariation among different indicators (Brown, 2006). However, unlike EFA where indicators are free to load on all factors, CFA requires that the model (factors and indicators) being tested be determined a priori. In other words, CFA requires a strong conceptual or empirical foundation to specify and evaluate a factor model. Another important distinction between EFA and

CFA is the ability of CFA models to correlate errors or unique variances. In EFA, the model must be specified with the assumption that measurement error is random; however, in CFA, the relationships may be freely estimated (Brown, 2006). Moreover, constraining all factors or all unique variances to be equal, is possible within CFA and can be useful for comparing different factor models (Brown, 2006).

Although not a requirement, running CFA after EFA can be useful because it provides a tentative identification of an underlying structure (Brown, 2006; Kline, 2016). As indicated previously, the questions in this measure were constructed through literature reviews and expert consensus, but the number of factors were hypothesized based on researcher judgment. In this study, the factor structure identified in EFA, with strong consideration for researcher judgement and theory, served as the underlying foundation for CFA. With those considerations in mind, a three-factor model is proposed to produce a good-fitting model for the BMIQ.

Figure 2

Hypothesized CFA Model



Scaling the Factors

Once the CFA model is identified, a second decision to be made relates to the metric of the factors. In CFA, regardless of the complexity of a model, the latent variables must be scaled to one of two things: the specifying marker indicators or fixing the variance to a value of 1.0 (Brown, 2006). The first method is the most common and produces an unstandardized solution, standardized solutions, and completely standardized solutions (Brown, 2006); this was the method used for this study. This is also the standard setting in the R statistical software used for study analyses (Macintosh Version 4.1.3; R High Sierra, 2021) package *Lavaan.*

Estimation Procedures

Following the metric factor decisions is selecting the appropriate estimation procedure. In CFA literature, maximum likelihood (ML) estimation is the most frequently used estimation procedure (Brown, 2006). ML evaluates the likelihood that the data (or observed covariances) were drawn from a population where the observed factor structure underlies the scores on the variables (Kline, 2016). ML assumptions include that the sample being used is large, has continuous indicators, and that the data is normally distributed (Brown, 2006). However, given the nature of the BMIQ and hypothesized non-normality, alternative estimations methods that are robust to non-normality were explored. The existing estimators with statistical corrections to standard errors and chi-square statistics include robust maximum likelihood (MLR) and diagonally weighted least squares (WLSMV). They have been suggested to be superior to ML when ordinal or non-normative data are analyzed (Li, 2016). This study used an MLR estimation given its tendency to outperform WLSMV in the small samples with non-normality.

Model Identification

Once the data have been prepared for analysis, the statistical identification of the model follows. Statistical model identification is important for the estimation parameters (i.e., the possibility to obtain a unique set of parameter estimates for each parameter in the model of unknown values; Brown, 2006). The parameters of the CFA model can only be estimated if the number of freely estimated parameters does not exceed the number of pieces of information in the input variance-covariance matrix (Brown, 2006). After estimation, a model can either be under-identified; that is, the number of unknowns (i.e.,

freely estimated parameters) exceeds the number of pieces of known information. A model can be just-identified when the number of unknowns is equal to number of knowns. In this case, a single set of parameter estimates perfectly fit the data. A model can be over-identified when the number of knowns exceeds the number of freely estimated model parameters. This model is the optimal model for goodness-of-fit statistics because the degrees of freedom (*df*; which are the difference in the number of knowns and the number of unknowns) are positive (Brown, 2006; Kline, 2016). Just-identified models have 0 *df* (because the number of knowns and unknowns are equal) and under-identified models have negative *df* (fewer unknowns than knowns; unsolvable). In the CFA for the BMIQ, there are 149 *df*. Therefore, in regards to degrees of freedom, the model is over-identified, as hoped for.

Having a model with positive degrees of freedom is a critical but an insufficient condition for model identification (Brown, 2006). Model identification can be susceptible to statistical under-identification as well as empirical under-identification. In empirically under-identified models, some aspect of the input matrix prevents the analysis from obtaining a valid set of parameter estimates. Examples include when all covariances equal 0. Usually, if an attempt is made to fit an empirically under-identified model, the computer software fails to yield a solution or provides an improper fit accompanied by error messages and warnings (e.g., *Heywood cases;* Brown, 2006). For this reason, all outputs for the CFA were screened for these potential issues.

Model Fit

Once the model is identified, it can be fit to the data. Three considerations for model fit are: (a) goodness-of-fit indices; (b) the presence or absence of localized areas of strain; and (c) interpretability, size, and statistical significance of the parameter estimates (Brown, 2006).

Goodness-of-fit Indices. Regarding goodness-of-fit indices, these tend to fall within three categories which include absolute fit, fit adjusting for model parsimony, and comparative or incremental fit (Brown, 2006). Absolute fit indices assess the overall model fit at an absolute level. The "classic" goodness-of-fit index is chi-square ($\chi 2$). In this index, the null hypothesis is tested so that a significant statistic supports the alternate hypothesis, meaning the model does not appropriately fit the data. Although popular, many criticisms have arisen in the literature for the fit index including how stringent it is and its extreme sensitivity to small and large sample sizes (Brown, 2006; Kline, 2016). Another absolute fit index is the Standardized Root Mean Square Residual (SRMR) which looks at the average discrepancy between the correlations observed in the input matrix and the correlation predicted by the model and is generally the preferred absolute fit index (Brown, 2006). SRMR can be values between 0.0 and 1.0, with 0.0 indicating "perfect fit" while values greater or equal to 0.10 generally indicate "poor fit" (Brown, 2006; Kline, 2016).

Meanwhile, parsimony correction indices incorporate a penalty for poor model parsimony. Root Mean Square Error of Approximation (RMSEA) is one of the most widely used and recommended indices within this category (Brown, 2006). It is a

population-based index that relies on evaluating reasonable fit versus whether the model holds absolutely (i.e., is equal as in the absolute statistics). The correction for parsimony occurs because it evaluates the discrepancy in fit per each degree of freedom (Brown, 2006). As with SRMR, RMSEA indicates a "perfect fit" when the values are 0.0 and a "poor fit" when values are greater or equal to 0.10 (Brown, 2006; Kline, 2016).

Finally, comparative fit indices evaluate the fit of a specified solution in relation to a more restricted, nested baseline ("null") model in which the covariances among indicators are fixed to 0 (Brown, 2006). The Comparative Fit Index (CFI) compares the amount of departure from close fit for the researcher's model against the baseline (null) model (Kline, 2016). CFI is the most popular index because, unlike the Tucker-Lewis Index (another comparative fit index), it is much easier to interpret due to the normed values between 0.0. and 1.0. (Brown, 2006; Kline, 2016). For CFI, values closer to 1.0 indicate a good fitting model.

For the current study, CFA model fit adequacy was based on several goodness-of-fit indices, which are RMSEA, CFI, and SRMR. These indices were selected based on their tendency to perform well regarding detecting model misspecification and, in some cases, the lack of dependence on sample size (Brown, 2006; Kline, 2016). Regarding assessing goodness-of-fit, for the current study, criteria were based on the general requirements for a "good fit model" rather than "best fit model" as generally approximated in Kline (2016) and Brown (2006). Thus, across these different indices, good fit was defined as having an RMSEA fit statistic <0.10, CFI ≥ 0.9, and SRMR <0.10. More stringent criteria have been recommended by simulation studies such as Hu

& Bentler (1999; e.g., RMSEA close to 0.06, SRMR close to 0.08, CFI close to 0.95) while others have criticized the use of goodness-of-fit indices generally, without consideration for other aspects of the analytic situation (Brown, 2006). Criticisms of this approach include: (a) the lack of consensus on cut-off criteria; and (b) how often indices are differentially affected by sample size, model complexity, estimation method used, normality of the data, amount and type of misspecification, and type of data (Brown, 2006). As such, while model evaluation began with the assessment of the goodness-of-fit indices and model characteristics (e.g., sample size), this study also examined a solution in terms of potential area of localized strain (e.g., Are there specific relationships that are not adequately reproduced?; Brown, 2006). To assess potential area of localized strain, two evaluations are commonly employed: residuals and modification indices (Brown, 2006).

Areas of Strain. The residual matrix provides information about how well the variances fit the covariances produced by the model's parameter estimates (Brown, 2006). One residual exists for each pair of indicators, and the larger the estimate, the poorer the fit (Brown, 2006; Kline, 2016). For interpretability, standardized residuals (analogous to z scores) are primarily used and were evaluated in the current study. Standardized residuals can either be positive or negative. Positive residuals indicate that the model underestimated the relations between two indicators, while negative standardized residuals indicate an overestimation to some extent (Brown, 2006). Because standardized residuals can be roughly interpreted as z scores, the z score values that correspond to conventional statistical significance levels are often employed as practical

cutoffs for standardized residuals. For example, a value of 1.96 is often used because it corresponds to a statistically significant z-score at $p < 0.5$; however, larger cut-offs are often recommended because of the potential influence by sample size (Brown, 2006). Thus, as per recommendations, a value of $z \geq 2.58$ was used in the current study to evaluate standardized residuals with outlying values.

Another aspect of model evaluation is the modification index (Brown, 2006). The modification index approximates how much the overall model $\chi 2$ would decrease if the fixed or constrained parameter were freely estimated. Generally, a good-fitting model should produce small modification indices, depending on the sample size. To address the sample size concern, an expected parameter changed (EPC) value is also considered. This statistic, which typically focuses on the completely standardized value, provides estimates of the expected value change (positive or negative) if the parameter were freely estimated (Brown, 2006). This statistic can be understood in a similar fashion to effect sizes (i.e., Cohen's *d*). However, such parameters should not be freed with the sole intent of improving model fit and instead must be justified based on prior research or theory. To further understand the model and evaluate it, this study discussed the modification index and standardized EPC. Adjustments to the model based on residuals or modification indices were considered when supported empirically.

Interpretability, Size, and Statistical Significance. A final aspect for the evaluation of model fit includes the interpretability, size, and statistical significance of the parameter estimates. This includes evaluating the direction, magnitude, and significance of the parameter estimates (i.e., the factor loadings, factor variances and

covariance, and indicator errors; Brown, 2006). The first step in this process is to screen for any error variances (e.g., Heywood cases), out-of-range values (e.g., standardized factor correlations exceeding 1.0), and negative factor variances (assuming the negative correlation is not in accordance with theory). The second step is to evaluate whether a parameter is necessary by evaluating its statistical significance. For instance, a nonsignificant estimate would indicate that eliminating the parameter from the model would not result in a significant decrease in the fit of the model and should be deleted. Third, it is important to evaluate the standard error of a parameter estimate to determine whether the magnitude of an estimate is appropriate. It is problematic if the standard error of the parameter estimate is too large or too small because it may indicate imprecise parameter estimates or the significance of the parameter would be difficult to calculate, respectively. Finally, it is important to determine whether the factor loadings in the model result in an estimate with a substantively meaningful magnitude. For instance, in applied research of questionnaires, completely standardized factor loadings of 0.30 and above are generally considered "salient" factor loadings or cross-loadings (Brown, 2006). Additionally, if factor loadings approach 1.0 (e.g., greater than 0.90), this provides strong evidence that factors may not be a distinct construct; thus, it is recommended to collapse factors or eliminate one (Brown, 2006).

Summary

In sum, procedural steps for CFA include: (a) model specification (e.g., indicators, factors); (b) input data (e.g., sample size); (c) model estimation (e.g., estimator used [MLR]); and (d) model evaluation (e.g., fit indices). Criteria for evaluating CFA are:

(a) lack of errors (e.g., Heywood cases); (b) RMSEA fit statistic <0.10, CFI ≥0.9, and SRMR <0.10; (c) standardized residuals ≥ 2.58; (d) consideration of the modification index and standardized EPC; (e) statistically significant parameters; (f) standardized factor loadings of 0.30 and above; (g) review of loadings greater than 0.90.

Reliability

To address the fourth research question, which is to determine the reliability of the confirmed scales, internal consistency reliability was evaluated. Reliability is a critical component to assess the utility of a measure. Reliability refers to the degree of precision or consistency exhibited when a measurement is repeated under identical conditions; that is, the overall proportion of true score variance to total observed variance (Boateng et al., 2018; Brown, 2006). A number of standard statistics have been developed to assess the reliability of a scale including test-retest reliability, split-half estimate, alternate-forms, and Cronbach's alpha (Boateng et al., 2018). Of these, Cronbach's alpha is one of the predominantly used methods to assess internal consistency reliability of scales (Raykov & Marcoulides, 2011).

Cronbach's alpha evaluates the internal consistency of the scale items (i.e., homogeneity of the items; Boateng, 2018). That is, when items that are believed to measure the same latent variable are strongly correlated, it is assumed that the test is reliable. On the other hand, if internal consistency is low, then the content of the items may be so heterogenous that the total score of the measure is not the best possible unit of analysis (Kline, 2016). Generally, when using Cronbach's alpha, acceptable coefficients

are greater or equal to 0.70 (Watkins, 2018). For the purposes of this study, acceptable coefficients were set at greater or equal to 0.70.

Results

Data was downloaded from Qualtrics into a master data file on Excel (version 16.6; Excel, 2022) that contained the BMIQ data and demographic information. The resulting database was de-identified and participant data were split in half (i.e., randomly assigned first two ratings and second two ratings for each user) to produce separate samples for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) procedures. In total, 184 evaluations of example BIPs using the BMIQ were completed by 47 participants. Missing data was deleted listwise before the analysis, which resulted in 88 completed BMIQ ratings for EFA and 88 completed BMIQ ratings for CFA. Analyses were conducted with the R statistical software (Macintosh Version 4.1.3; R High Sierra, 2021) and several of its packages (i.e., *psych, graphics, GPA rotation, haven, dplyr, laavan, semPlot*) for EFA and CFA.

Preliminary Factor Structure Identification

The goal of the first research question was to identify dimensions (i.e., factors) of intervention quality that emerge from a preliminary pool of BMIQ items using EFA. Before beginning the factor extraction procedures, several criteria were explored to ensure that the data were factorable. Univariate skewness and kurtosis were analyzed first and were observed to not be extreme (see Table 4). Then, a correlation matrix was examined to identify any low or high inter-item correlations (see Table 5). No items demonstrated multicollinearity (i.e., high inter-item correlations with at least three other

items in the scale); therefore, no items were deleted due to high inter-correlations. However, Item 16 had multiple inter-item correlations less than 0.30; thus Item 16 was deleted (Fabrigar et al., 1999; Hair et al., 2010; Watkins, 2018). Finally, the KMO statistic and Bartlett's test of sphericity were assessed. The KMO statistic for the overall correlation was 0.90 and all remaining items were above 0.80, well over the minimum standard for conducting factor analysis. The results of Bartlett's test of sphericity indicated that the correlation matrix was significant, $\chi 2(231) = 1811$, $p < .0001$. Based on these indicators, the matrix was deemed appropriate for factor analysis. Bartlett's test indicates that the correlation matrix was not an identity matrix, and the sample size was found to be sufficient relative to the number of items remaining. Descriptive statistics for the data set are provided in Table 4.

Table 4

Item-level Descriptive Statistics by EFA and CFA Samples

	EFA				CFA			
Item	Mean	*SD*	Skew	Kurtosis	Mean	*SD*	Skew	Kurtosis
Q1	3.85	1.29	-1.27	1.12	3.86	1.30	-1.27	1.10
Q2	3.80	1.39	-1.28	1.03	3.74	1.31	-0.93	0.10
Q3	3.80	1.13	-0.58	-0.57	3.78	1.20	-1.16	1.38
Q4	3.62	1.46	-1.10	0.39	3.60	1.37	-1.01	0.40
Q5	3.97	1.19	-1.22	1.26	3.79	1.23	-1.09	0.74
Q6	4.07	1.04	-1.29	1.84	3.86	1.20	-1.17	1.17
Q7	3.98	1.11	-1.06	0.98	3.84	1.30	-1.05	0.39
Q8	3.97	1.14	-1.16	0.94	3.79	1.33	-1.15	0.68
Q9	3.92	1.06	-1.29	2.40	3.67	1.25	-0.93	0.32
Q10	3.89	0.97	-0.91	1.45	3.70	1.23	-1.02	0.62
Q11	3.55	1.29	-0.80	0.08	3.46	1.35	-0.72	-0.13
Q12	3.59	1.37	-0.99	0.28	3.56	1.41	-0.92	0.06
Q13	3.84	1.08	-0.51	-0.62	3.75	1.15	-0.76	0.10
Q14	3.95	0.96	-0.60	-0.30	3.93	1.10	-1.07	1.05
Q15	3.83	1.14	-0.77	-0.24	3.84	1.11	-0.85	0.40
Q16	2.47	1.73	-0.05	-1.38	2.62	1.66	-0.27	-1.18
Q17	3.84	1.06	-0.68	-0.08	3.76	1.09	-0.64	-0.34
Q18	3.83	1.07	-0.67	-0.38	3.79	1.11	-0.95	0.62
Q19	4.16	0.97	-1.36	2.47	3.83	1.24	-1.01	0.29
Q20	3.95	1.01	-1.05	1.34	3.67	1.34	-1.04	0.39
Q21	3.94	1.14	-1.31	1.89	3.64	1.26	-0.81	-0.14
Q22	3.91	`.14	-1.28	1.85	3.83	1.18	-1.04	0.72
Q23	3.99	1.10	-1.31	2.01	3.95	1.08	-1.23	2.14

Note. EFA = exploratory factor analysis. CFA = confirmatory factor analysis.

Table 5

Item Correlations

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22
Q2	.80																					
Q3	.65	.64																				
Q4	.77	.86	.70																			
Q5	.60	.70	.77	.70																		
Q6	.56	.48	.50	.54	.62																	
Q7	.44	.32	.40	.38	.44	.81																
Q8	.48	.41	.47	.38	.47	.75	.74															
Q9	.71	.72	.68	.73	.64	.57	.47	.60														
Q10	.66	.64	.65	.70	.57	.58	.51	.59	.78													
Q11	.61	.64	.58	.65	.54	.50	.37	.48	.82	.74												
Q12	.74	.76	.63	.78	.62	.56	.38	.42	.80	.72	.85											
Q13	.37	.36	.50	.38	.50	.43	.34	.39	.38	.39	.40	.38										
Q14	.45	.31	.57	.40	.52	.51	.48	.55	.43	.39	.36	.40	.68									
Q15	.38	.28	.52	.31	.48	.54	.49	.57	.39	.42	.36	.33	.58	.69								
Q16	.02	-.01	.01	.01	.04	.02	-.02	-.06	.06	.08	-.09	-.09	.08	-.05	-.16							
Q17	.56	.51	.56	.62	.47	.53	.40	.50	.61	.60	.61	.60	.65	.56	.36	.05						
Q18	.45	.40	.54	.49	.43	.52	.48	.45	.47	.52	.43	.50	.66	.66	.51	-.01	.73					
Q19	.54	.56	.57	.54	.58	.56	.48	.51	.56	.52	.48	.55	.62	.53	.47	.06	.57	.70				
Q20	.47	.54	.64	.63	.60	.42	.32	.38	.46	.46	.38	.48	.46	.50	.45	-.09	.49	.49	.46			
Q21	.53	.49	.65	.54	.63	.58	.38	.47	.61	.59	.57	.62	.47	.59	.44	.05	.55	.55	.60	.56		
Q22	.49	.56	.64	.60	.61	.53	.43	.48	.57	.57	.52	.51	.47	.57	.43	.01	.58	.55	.64	.68	.73	
Q23	.43	.57	.61	.61	.52	.41	.33	.37	.56	.47	.55	.54	.50	.59	.33	-.05	.62	.49	.50	.70	.60	.79

Note. All correlations are significant at the 0.01 level.

The parallel analysis and scree plot can be found in Figure 3 and Figure 4, respectively. Parallel analysis, MAP (minimum = 0.04), and scree all suggested that three factors should be retained.

Figure 3

Scree Plot

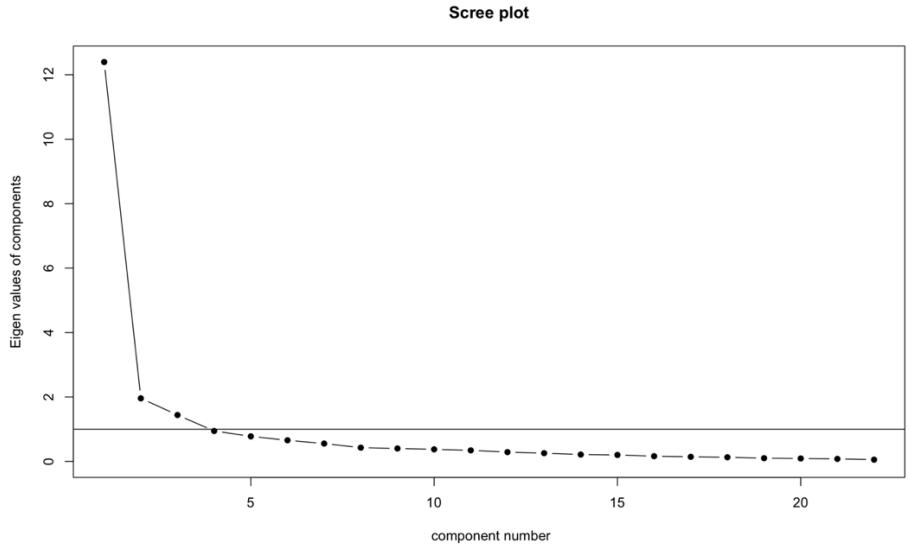
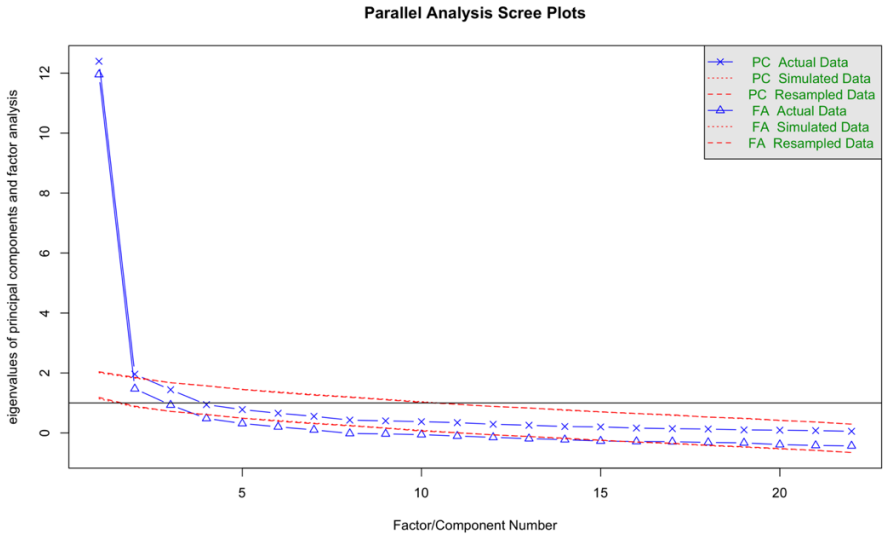


Figure 4

Parallel Analysis Plot



Following item deletion and retention (discussed in the following section), the EFA revealed that Factor 1 accounted for 30% of the variance, Factor 2 accounted for 25% of the variance, and Factor 3 accounted for 13% of the variance in the items. In total, the extracted factors accounted for 69% of the total variance. Factor 1 and Factor 2 correlated at 0.63, Factor 2 and Factor 3 correlated at 0.50, and Factor 1 and Factor 3 correlated at 0.43, which indicates that the subscales are distinct and below the a priori established threshold. Given these results, the three-factor solution was accepted as adequate.

Item Retention

The second research question focused on identifying item combinations that most efficiently and appropriately capture intervention quality (i.e., item retention/removal). The three-factor solution was examined for adequacy. In the three-factor solution identified during the EFA, each factor was observed to be saliently loaded by at least three items. However, several items were observed to cross load on two factors: Items 3, 5, 15, 17, 21, 23. These items were subsequently deleted one by one at each repeat of the analyses. Items 3, 5, and 15 were deleted. The additional items (i.e., Item 17, 21, 23) were no longer observed to cross load after Item 15 was deleted. Following the final item elimination and EFA repeat, evaluation of the communalities across the BMIQ items revealed that the items ranged between 0.54 and 0.82 so all remaining items were retained. Pattern coefficients revealed loadings between 0.52 and 0.89, which are moderate to strong, and all were retained. See Table 6 for the communality and pattern coefficients for each item.

Table 6

Item-level Pattern Coefficients from Exploratory Factor Analysis

Item	Item Stem	Factor 1	Factor 2	Factor 3	h^2
Q1	The intervention clearly connects to a hypothesized function maintaining or creating the problem (i.e., seek/obtain, escape/avoid - acquisition, fluency, generalization).	**0.74**	0.01	0.14	0.67
Q2	Overall, the intervention procedures appear to be appropriate for the presenting concern(s).	**0.89**	0.04	-0.09	0.77
Q4	The match between intervention mechanisms and desired outcomes is intuitive (i.e., the intervention appears likely to improve the targeted student functioning/behavior).	**0.83**	0.15	-0.09	0.80
Q6	Implementation procedures are clearly presented.	0.21	0.12	**0.70**	0.79
Q7	Implementation procedures are thoroughly explained.	0.03	0.05	**0.83**	0.75
Q8	The role of the interventionist is clearly defined.	0.12	0.10	**0.73**	0.72

Q9	The intervention appears grounded in easily identified or recognized theoretical perspective (e.g., behaviorism, cognitive-behavioral, direct instruction)	**0.81**	-0.02	0.20	0.81
Q10	The intervention appears grounded in a widely accepted theoretical perspective (e.g., behaviorism, cognitive-behavioral, direct instruction).	**0.68**	0.02	0.26	0.70
Q11	The intervention is clearly supported by a solid evidence-based (i.e., research).	**0.79**	0.01	0.09	0.69
Q12	The intervention is consistent with best-practice recommendations for addressing the problem/difficulty of concern.	**0.89**	0.00	0.03	0.82
Q13	Intervention implementation relies on an accessible and reasonable amount of human resources (i.e., staffing).	-0.14	**0.82**	0.05	0.59
Q14	Adequate resources (e.g., equipment, materials, notes) to support the implementation are available and accessible for the duration of the intervention	-0.18	**0.77**	0.23	0.65

Q17	Overall, the intervention appears to be one that an interventionist/teacher/parent/educator would be willing to use.	0.25	**0.58**	0.07	0.63
Q18	The intervention appears implementable for most educators.	-0.04	**0.74**	0.16	0.65
Q19	The intervention is appropriate for the intended setting/context (e.g., school, home).	0.17	**0.55**	0.17	0.59
Q20	The intervention is appropriate for the intended student/client (e.g., age, needs, disability, eligibility, background).	0.24	**0.62**	-0.14	0.54
Q21	The intervention is appropriately sensitive/adaptive/responsive to the student/client's disability/ability status.	0.28	**0.52**	0.07	0.59
Q22	The intervention is appropriately sensitive/adaptive/responsive to the student/client's language or communication skills and abilities.	0.22	**0.68**	-0.02	0.67
Q23	The intervention is appropriately sensitive/adaptive/responsive to the student/client's racial/ethnic background.	0.29	**0.68**	-0.20	0.64

Note. Significant (i.e., above 0.30), retained pattern coefficients noted in bold; h^2 = communalities.

Reliability Estimates for Hypothesized Scales

EFA revealed a three-factor model. Factor 1 contained seven items that measure whether it is *Sensible* to use the intervention. This refers to whether the intervention is supported by theoretical or empirical evidence and whether it functionally addresses the problem. The coefficient alpha was found to be acceptable ($\propto = 0.95$; CI = .93-.97). Factor 2 contained nine items that measure how *Acceptable* the intervention might be. That is, whether the intervention procedures and components are feasible, manageable (i.e., implementers can implement the intervention as designed), and reasonable to use with clients. This factor was found to be acceptable ($\propto = 0.93$; CI = .90-.95). Finally, Factor 3 contained three items that measure how *Procedurally Clear* the intervention is. In other words, the items evaluate the precision, quality, and clarity with which all intervention components and procedures are communicated. Factor 3 was found to be acceptable with an alpha value of 0.91 (CI = .87-.94). The mean scores for each factor were: Factor 1 $M = 3.75$ ($SD = 1.11$), Factor 2 $M = 3.94$ ($SD = 0.84$), Factor 3 $M = 4.00$ ($SD = 1.00$).

Factor Structure Confirmation

The third research question sought to confirm the initially identified three-factor model (via EFA) in a secondary sample of BMIQ ratings. CFA was used to address this research question. Prior to evaluating the initially identified three-factor model, a general, single-factor model was analyzed. Initial screening of the single-factor CFA output did not display any error messages or out-of-range values that may have highlighted an empirical under-identification. Goodness-of-fit indices were reviewed. Using the a priori

guidelines, the single-factor model demonstrated poor fit (RMSEA=.16, CFI=.75, SRMR=.09) suggesting that the items in the BMIQ do not measure a unidimensional construct (Brown, 2006).

Next, the three-factor model that emerged from the EFA (i.e., Sensible, Acceptable, Procedurally Clear) was analyzed. Initial screening of the output did not display any error messages or out-of-range values. Goodness-of-fit indices were reviewed. The three-factor model, using an MLR estimation, was deemed to be a good fit using CFI (0.90) and SRMR (0.06) based on the a priori guidelines; however, not all criteria were met (RMSEA = 0.10). See Table 7 for a comparison of the models. Inspection of the standardized residuals covariance matrix indicated there were no areas of strain. The largest standardized residual was $z = 0.22$. Modification indices were examined to review potential improvements to the three-factor model; however, the largest modification indices were not theoretically justified in the hypothesized model.

Table 7

Goodness-of-Fit Statistics

Model	X^2	*df*	RMSEA	CFI	SRMR
1	478.44	152	0.16 (.14-.17)	0.75	0.09
2	288.49	149	0.10 (.09-.12)	0.90	0.06

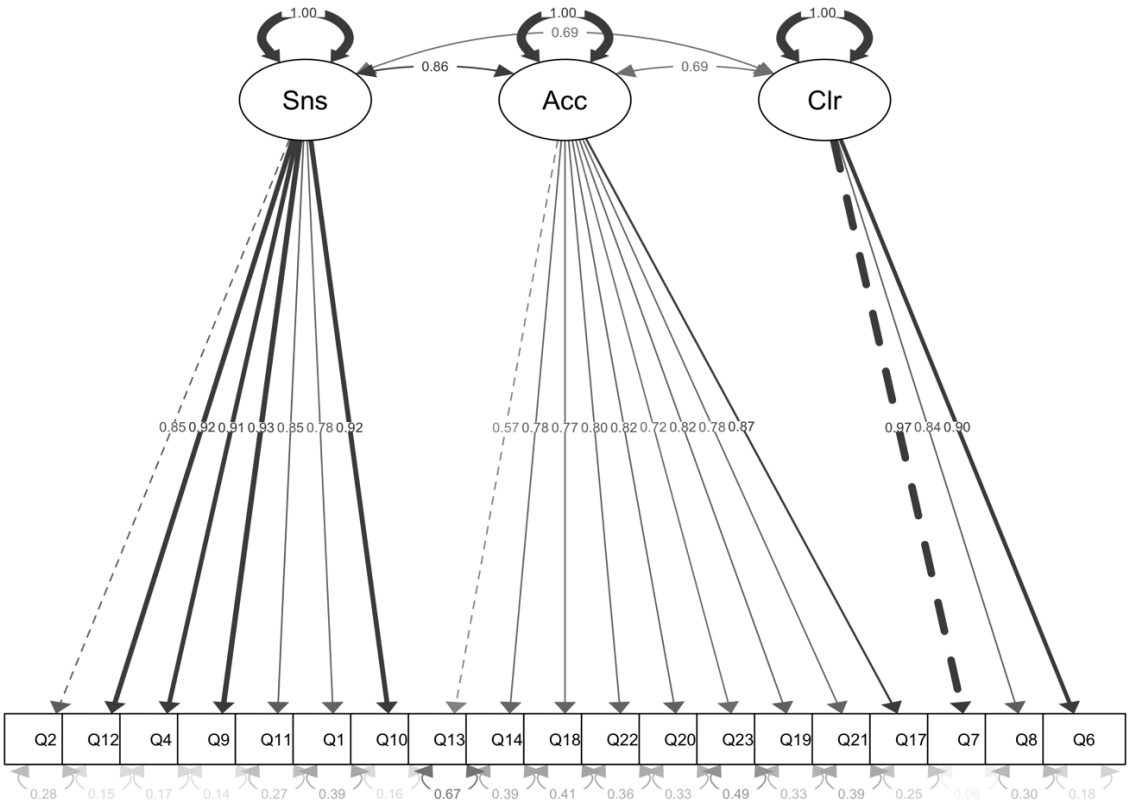
Note. df=Degrees of Freedom. CFI = Comparative Fit Index. RMSEA = Root Mean Square of Error of Approximation. SRMR = Standard Root Mean-Squared Residual.

Inspection of the factor loadings for this model indicated that there were no insignificant factor loadings (i.e., <0.30; see Figure 5). All indicators loaded on their

respective factors significantly and loadings ranged from moderate (0.57) to high (0.97). The correlations between the factors ranges from moderate (0.69) to large (0.86). No factors were correlated $r \geq .90$, indicating a lack of multicollinearity. The lack of multicollinearity and the presence of medium to large correlations points to the factors being related, but distinct and different constructs as measured by the items.

Figure 5

CFA Three Factor Model



Reliability Estimates for Retained Scales

The fourth research question sought to evaluate levels of reliability within retained BMIQ scales. To address this research question, internal consistency estimates

were calculated for each retained BMIQ factor. The *Sensible* factor contains the same seven items as in the EFA. The coefficient alpha was found to be acceptable ($\propto = 0.96$; CI = .95-.97. The *Acceptable* factor contains the same nine items and was found to be acceptable ($\propto = 0.93$; CI = .90-.95). Finally, the *Procedurally Clear* factor continues to have three items and was found to be acceptable with an alpha value of 0.93 (CI = .90-.95). The mean scores for each factor were: Sensible $M = 3.66$ ($SD = 1.18$), Acceptable $M = 3.80$ ($SD = 0.94$), Procedurally Clear $M = 3.83$ ($SD = 1.19$).

CHAPTER FIVE: DISCUSSION OF FINDINGS, LIMITATIONS, AND FUTURE DIRECTIONS

Discussion

Having an effective consultation process is a necessary component of developing or selecting an appropriate, evidence-based intervention within an indirect service delivery model. However, a good problem-solving consultation process does not guarantee a good, quality intervention because of the multiple interdependent components involved in school consultation. A breakdown in any of the aspects involved could result in negative or null outcomes; thus, making it difficult to differentiate the quality of the consultative process from the quality of the intervention being implemented. The BMIQ was developed to assess intervention quality. More specifically, the BMIQ was developed to address the potentially confounding impact of intervention variability within consultation research and for the vetting and adoption process of interventions in applied school settings. This may offer a solution for improving the internal validity of research evaluating the effectiveness of school consultation. Despite the importance to do so, no clear guidelines have been established to define the characteristics of high-quality interventions within this context. For the current study, quality interventions were defined as those with being culturally sensitive/adaptive/appropriate (i.e., the intervention accounts for the socio-cultural values, beliefs, norms, and identity characteristics of the student/client), functionally matched (i.e., the intervention addresses the underlying cause of the problem), procedurally clear (i.e., the precision, quality, and clarity with which all intervention

components and procedures are communicated), evidence-based (i.e., the connection of the intervention to theoretical and/or empirical evidence supporting its use), economical (i.e., the amount of time and resources [money, materials] necessary for implementation, and acceptable (i.e., the degree to which intervention procedures and components are feasible, reasonable, and manageable [implementers can implement the intervention as designed]).

To begin accumulating validation evidence to support the hypothesized constructs underlying the BMIQ and to support continued scholarly efforts, four research questions were addressed: (a) What factors emerge for the BMIQ? (b) What are the appropriate number of items to retain to maximize findings? (c) Does the factor structure identified hold true in further analyses? (d) Is the BMIQ a reliable tool? What follows is a discussion of the results of this study. The chapter will conclude with a consideration of study limitations and future research directions.

Preliminary Content Validation

Prior to formal, empirical analyses of the BMIQ, a preliminary content validation activity was conducted. Content validity in conjunction with literature review were used to inform preliminary domain development, inform preliminary item-domain assignment, initial pruning of unnecessary or repetitive items, and justification of the remaining items. Goals of this activity were to accumulate validity evidence to address scoring inferences (i.e., rules and procedures for obtaining the observed score; Brennan, 2013). More specifically, the BMIQ scoring inference would postulate that constructed items relate to the underlying elements or domains of intervention quality as well as intervention quality

as a broader construct. This activity facilitated the accumulation of information supporting the use of developed items to this end (i.e., evaluating intervention quality). Additionally, these findings address extrapolation inferences related to the proposed BMIQ interpretations and uses. Validity evidence addressing extrapolation inferences, or the connection of performance scores to real-world performance (Cook et al., 2013) is generated using methods that ensure that the assessment reflects the key aspects of real performance and empiric analyses evaluating the relationship between the test performance and real-world performance (Brennan, 2013). Item prioritization, retention, or elimination based on response from individuals with knowledge, training and experience in the real-world tasks targeted by the BMIQ (i.e., assessment, consultation, intervention identification/development) support the assertion that test domain(s) reflects the key aspects of real performance.

Preliminary Factor Structure Identification

The goal of the first research question was to identify dimensions (i.e., factors) of intervention quality that emerge from a preliminary pool of BMIQ items. Identification of the underlying factor structure served two broad purposes. First, it served to identify dimensions or factors to facilitate the refinement of the BMIQ, specifically the empirically driven reorganization of items under the identified factors. Second, addressing this research question begins the accumulation of validation evidence supporting the extrapolation inferences underlying the BMIQ interpretations and uses. To address this research question, EFA was conducted using 88 completed BMIQ ratings from 47 participants following listwise deletion of missing data. Evaluation of the data

determined it was factorable. Using a PA method of extraction with an oblimin rotation extracted three factors, which included the constructs Sensible, Acceptable, and Procedurally Clear. Of the preliminary BMIQ factors that were hypothesized based on a review of existing intervention quality literature, only Procedurally Clear was retained based on the findings of EFA. The other five categories (i.e., Culturally Sensitive/Adaptive/Appropriate, Functionally Matched, Evidence-Based, Economical, Acceptable) required transformation. More specifically, Economical, Acceptable, and Culturally Sensitive/Adaptive/Appropriate were combined into one category, which was renamed "Acceptable." Items in this newly formed factor targeted the appropriateness of the intervention to the client (e.g., skills, background, needs), appropriateness to the context, and adequacy (e.g., resources, staff). Similarly, Functionally Matched and Evidence-Based were combined and renamed "Sensible." Items in this newly formed factor targeted whether the intervention addressed the function of the behaviors of concern, was appropriate for the presenting problem, based on evidence, and consistent with the research based for addressing the presenting problem.

When compared to existing literature, it is not surprising that ultimately, initial hypothesized categories required some refinement. Previous empirical research has not clearly separated some of the hypothesized constructs (e.g., Functionally Matched, Evidence-Based) as separate constructs. For example, Sanetti et al. (2014), discussed the importance of matching an evidence-based intervention to the function of the presenting problem, suggesting they may be interrelated to some degree. Further research and a larger sample size is needed to determine whether these two constructs are related but

separate measures of intervention quality. On the other hand, it is surprising that the hypothesized constructs Economical, Acceptable, and Culturally Sensitive/Adaptive/Appropriate were not identified as separate constructs. For example, previous empirical research has identified these variables as distinct mediators/moderators of treatment fidelity (e.g., Sanetti & Kratochwill, 2009). One explanation is that the lack of identification as separate constructs in this study may be due to the small sample size. Additional ratings may have shown these to be distinct, critical constructs. Another explanation is that these variables as separate entities are not good descriptors of intervention quality, and as suggested by the current analyses, may in fact be a single construct. However, an alternative explanation may be how the items measured each construct. It may be that questions were much better indicators of Acceptability than they were of the Economical or Culturally Sensitive/Adaptive/Appropriate constructs. For instance, when compared to other measures of acceptability such as the BIRS (Von Brock & Elliott, 1987), there is some overlap between items from the BMIQ regarding acceptability and cultural sensitivity/appropriateness, and the BIRS (i.e., "This intervention would be an appropriate intervention for a variety of children").

Findings from EFA also serve to begin the accumulation of validity evidence to address extrapolation inferences. As noted previously, extrapolation inferences involve moving from observed to real-world performance. This evidence takes two primary forms, activities that support beliefs that assessment reflects the key aspects of real performance and empiric evaluations of the relationship between the test performance

(e.g., domain specification, construct definition; Cook et al., 2015) and real-world performance (e.g., concurrent validity, predictive validity; Cook et al., 2015). The extrapolation inference serves to evaluate how well people can perform, or at least predict how they might perform, certain activities over some range of conditions (Kane, 2013; Cook et al., 2015). The results of EFA served to address the scope of test (e.g., domain specification, construct definition), doing so in a moderately authentic assessment context (i.e., case study intervention evaluated by educators, educators in training, educational support specialists).

Item Retention

The second research question focused on identifying item combinations that most efficiently and appropriately capture intervention quality (i.e., item retention/removal). Goals related to this research questions included: a) to refine the preliminarily hypothesized BMIQ items by reducing the number of items for each identified factor; and b) accumulate initial validation evidence to address scoring and generalization inferences related to the proposed BMIQ interpretations and uses. Some refinement was made to the BMIQ during initial development and content validation activities. Further refinements were made during the EFA. Ultimately, instrument refinement seeks to maximize the valid and reliable information it generates in the fewest, most efficiently completed number of items possible. The preliminary version of the BMIQ post validation trials contained 23 items and, following EFA analysis, the final BMIQ version was streamlined; it contained only 19 items. Four items were ultimately excluded (Item 3, 5, 15, 16). Item 3 (*The intervention is appropriate for the student's knowledge, skills, and*

abilities) and Item 5 (*The intervention is developmentally appropriate for the student/client*) were both from the hypothesized construct Functionally Matched. The remaining items, Item 1 (*The intervention clearly connects to a hypothesized function maintaining or creating the problem [i.e., seek/obtain, escape/avoid - acquisition, fluency, generalization]*), Item 2 (*Overall, the intervention procedures appear to be appropriate for the presenting concerns*), and Item 4 (*The match between intervention mechanisms and desired outcomes is intuitive [i.e., the intervention appears likely to improve the targeted student functioning/behavior]*) remained as part of the Sensible construct. After analyzing the wording of the remaining items, these items appear to better inquire about the match between the intervention and the client compared to the deleted items. An analysis of the deleted items suggests these questions may have been better suited to address the appropriateness of the intervention rather than functional match.

Another item that was deleted based on the EFA was Item 15 (*Intervention implementation relies on an accessible and reasonable amount of material resources [e.g., workbooks, flashcards, tangible rewards, technology]*). This item was from the Economical construct. Compared to the remaining items within this category (i.e., Item 13 [*Intervention implementation relies on an accessible and reasonable amount of human resources, i.e., staffing*]; Item 14 [*Adequate resources, e.g., equipment, materials, notes, to support the implementation are available and accessible for the duration of the intervention*]) Item 15 was very similarly worded to Item 14. The primary distinction between the two was the indication of duration in Item 14. It may be that having adequate

resources for the duration in which the intervention is implemented is more applicable to the situation than general access to the materials for an unspecified amount of time.

The final item eliminated from the BMIQ was Item 16 (*The intervention appears too complex/difficult to implement [e.g., numerous components or elements]*). This item belonged to the Acceptable construct and was the only negatively worded question. This item was deleted before the EFA analysis because of its poor correlation to other items in the data analysis stage (i.e., several correlations <0.30). Given this, this item may need to be revised so that it is positively worded like the other items in the BMIQ so a determination could be made about its importance and relevance to the measurement of quality. Additionally, it may be that this item is not a good indicator of intervention acceptability in general. Some measures have defined intervention complexity as a measure of feasibility rather than acceptability (e.g., URP-I), which the current study attempted to combine.

Item reduction using EFA results further the early accumulation of validity evidence related to the proposed interpretations and uses of the BMIQ. First, the scoring inference relates to the rules and procedures for obtaining the observes score(s) that are used for interpretations and decisions (Brennan, 2013). Item refinement or reduction directly relates to improving the manner (i.e., item content) and procedures (i.e., number of items completed) used by the BMIQ to generate scores. Furthermore, these results contribute evidence that the overall quality score produced by the BMIQ may be able to be used to determine overall intervention "quality" construct using the determined sets of items. Similarly, Brennan (2013) notes that inferences are often not as delineated as they

appear and scoring and generalization inferences may appear similar, or overlap. Generalization inferences reference the connection of observed scores to all possible scores, or the expected values of observed scores over investigator-specified conditions of measurement (Brennan, 2013). The two primary sources of validity evidence that address the generalization inference come from adequate and appropriate sampling and through evidence of reliability (Cook et al., 2015).

Factor Structure Confirmation

The goal of the third research question was to confirm the previously identified factor structure for the BMIQ. Confirmation of the underlying factor structure served two broad purposes. First, confirmation of the previously identified factor structure would increase the confidence in the use of the BMIQ to evaluate intervention quality across the dimensions noted. In short, this goal sought to further address extrapolation inferences related to the BMIQ. Second, if unable to confirm the initially identified factor, results would allow further refinement of the BMIQ. CFA results indicated the three factors were distinct (i.e., correlations between factors fell below 0.90) with excellent item loadings (i.e., greater than 0.30), suggesting there was sufficient discriminant validity to conclude that the factors measure independent dimensions of quality. After examining the combined evidence of model fit, factor loadings and factor inter-correlations, the model structure did not hold entirely true during the CFA analysis of this preliminary validation study. The three-factor model was found to be adequate using the CFI (0.90) and SRMR (0.06) fit indices based on their a priori cut-offs. On the other hand, the RMSEA value (0.10) suggested that the model fit was inadequate. One explanation for this is that

RMSEA values were impacted by the sample size. It has been suggested that when fit indices fall in these "marginal" ranges, it is vital to consider the consistency of model fit as expressed by the various types of fit indices in conjunction with the aspects of the particular situation (Brown, 2006). For example, when N is somewhat small, a poor fitting model based on an RMSEA >0.08 may be of less concern if all other indices are in a range suggesting "good" model fit (Brown, 2006). To further this point, Breivik and Olsson (2001; as cited in Kline, 2016), using simulation studies, found that the RMSEA tends to impose a harsher penalty on smaller models with relatively few variables; such is the case in the current study. This is because smaller models have fewer degrees of freedom, but larger models have more "room" for higher degrees of freedom values (Kline, 2016). In light of this evidence, the proposed three-factor model is likely demonstrating a fair fit to the preliminary data despite its small sample size.

Like EFA, CFA results contribute to early validity evidence addressing extrapolation inferences. Again, extrapolation inferences involve moving from observed to real-world performance (Kane, 2013; Cook et al., 2015). These findings provide additional evidence related to the scope of test (e.g., domain specification, construct definition) within a context that is similar to those anticipated for applied BMIQ use (i.e., case study intervention evaluated by educators, educators in training, educational support specialists).

Reliability Estimates for Retained Scales

The aim of the fourth research question was to evaluate levels of reliability within retained BMIQ scales. Evaluating reliability within BMIQ scales served to address

generalization inferences related to proposed BMIQ interpretation and uses by determining whether it delivered reliable scores. Internal consistency reliability was measured with Cronbach's alpha. Study findings suggest that internal consistency reliability estimates for all subscales exceeded desired levels (i.e., $\alpha \geq 0.90$). Because internal consistency reliability is high (>0.70; Watkins, 2018), this may suggest that the items within each subscale are appropriate for measuring their respective subscale (i.e., acceptability, sensibility, procedural clarity) accurately and consistently by the intended users (e.g., researchers). Additional validation evidence is necessary, but the nature of these results is promising given the noted levels of reliability.

These results begin the accumulation of validity evidence that address scoring inferences. Results, when used to inform item-domain alignment, influence the structure and procedures (i.e., item presented to users) of the BMIQ (i.e., scoring inference).

Limitations

As noted previously, the present preliminary validation study aimed to investigate and improve the BMIQ using a sample of students, faculty, and practitioners from School Psychology, Education, Special Education, and closely related fields using random assignment of these respondents to different BIPs. However, the design of the present study is not without limitations. First, this preliminary study has a small sample size (N = 47, observations = 88) which impacted the level of confidence in the current results. Although sample size recommendations can vary, at the very least, low estimates for structural equation modeling suggest that samples should include 100 observations per analysis (Brown, 2006). Other estimates include 10 observations for each number of

survey items, or upwards of 300 respondents. Guadagnoli and Veliver (1988) suggests a minimum of 300-450 for acceptable comparability of patterns when using factor analysis. Regardless of the cut-off selected, the current sample of 88 observations for each analysis is poor in comparison. This may have future implications for the data and development of the survey. For example, the current three factor model and deletion of four items was based on the current 88 observations for EFA and CFA. However, it may be that by increasing the sample size to at least 10 observations for each original item (e.g., 230 observations) may result in the emergence of an additional factor, may influence the merging of current factors, may change the number or type of items retained, or the reliability may deteriorate. In each case, the results of this preliminary validation study would not hold true.

Second, broadly are the limitations related to the questionnaire itself. The BMIQ was developed to assess the quality of an intervention within a consultation process. However, there are currently no clear guidelines as to what constitutes a "high quality" intervention. Previous empirical literature has suggested that good interventions are those that are based on evidence, viewed as acceptable, feasible, are clear to implement, culturally adaptive/sensitive/appropriate, targeted to the problem at hand, and are clear about the resources (e.g., time, money, training) needed (Domenech Rodríguez et al., 2011; Erchul & Martens, 2010; Gersten et al., 2008; Johnson et al., 2018; Molleman et al., 2006; Sanetti & Gritter, 2010; Sanetti et al. 2014). The quality intervention constructs developed for this measure were hypothesized based on this previous literature, but, as

with any novel measure, lacks the strong, explicit foundation (e.g., previous validity studies, previous measures) to support its hypothesized structure.

Future Directions

Future research is needed to continue the accumulation of validity evidence to refine the BMIQ. First, a larger sample with diverse participants (e.g., undergraduates, graduates, and established researchers) that meets factor analysis requirements is warranted to confirm or refine the results of the current study to provide additional evidence for the scoring and generalization inferences. A study with a larger sample of diverse participants will provide the opportunity to re-run the factor analysis to better assess and establish the correlations and inter-item correlations observed in this pilot study (e.g., cross-validate; i.e., generalization inference). A larger, more diverse sample will help determine whether the associated items continue to support the constructs they are designed to assess. For example, a larger diverse sample will help determine whether Sensible and Acceptable being distinct constructs remains true. In the current pilot study, the Sensible and Acceptable constructs in the CFA have a rather large correlation ($r = 0.86$). The current study's a priori decision for questionable factor correlations was 0.90; however, some researchers have a lower cut of 0.85 suggesting that some may view these factor correlations as indistinct (Brown, 2006). As such, these factors may be candidates for elimination (i.e., revert to a two-item factor) as they may be redundant. Another example is that a larger, more diverse sample will allow a close inspect of the items eliminated and retained in this study (i.e., scoring inference). After confirming in a follow up study that the current items and factors are retained, future studies may want to

evaluate whether the items not retain should be reworded or whether additional considerations are warranted. As just one example, in the health sciences, when programs are being evaluating, the necessary training of the staff implementing the programs is frequently considered (see Molleman et al., 2006). The BMIQ considered resources as material resources (e.g., workbooks) and human resources (e.g., adequate staff) rather than required training. For future iterations of this measure, it may be important to consider including a question regarding whether staff appear to have the necessary training to implement the intervention.

Second, following the recruitment of a larger and more diverse participant sample, future research should evaluate interrater reliability, which traditionally is used as a means to evaluate reproducibility and generalization (Cook et al., 2015). This will provide additional validation evidence that the scores of the BMIQ can be scored accurately and consistently by the intended users. Third, future research should evaluate the interpretation and use of the BMIQ with a variety of interventions developed withing a school consultative framework to garner evidence for the extrapolation inference. The current preliminary study included 21 BIPs (i.e., behavioral interventions) of varying quality to evaluate the interpretation and use of the BMIQ. Using varied interventions, such as academic-based interventions developed within a consultative process, will assist to begin generalizing the accuracy and applicability of the BMIQ to a range of interventions.

Fourth, to accumulate evidence for the implications inference, future empirical research should evaluate the difference in test performance among different subgroups

for which performance is expected to be similar (Cook et al., 2015). For example, among the recommended large group of diverse participants, it may be useful to compare novice school psychology researchers (e.g., Year 1-4 graduate researchers) to intermediate groups (Year 5+ or early professionals; i.e., differential item functioning). Should there be a difference in performance, it would allow for an evaluation and assessment of the BMIQ's intended or unintended consequences (e.g., training requirements). Fifth, future research should evaluate the impact of BMIQ on intervention adoption in practice; that is, if an intervention is deemed to be of a low-quality, how might that impact the later adoption of the intervention or overall perception of the consultation study in which in the intervention was used? This implications evidence may be useful for understanding additional intended or unintended consequences of the BMIQ.

Finally, while other "quality" intervention measures are limited, it may be useful for future studies to do an item comparison on the identified factors (e.g., acceptability) contained within the BMIQ to established measures (e.g., BIRS). This will serve as extrapolation evidence by providing concurrent validity evidence to an aligned measure (Kane, 2013).

Conclusions

School consultation has been shown to be both an effective and efficacious service delivery model. Several studies have evaluated both its efficacy and effectiveness across a broad range of situation; however, to date, few studies have attempted to discern the consultative problem-solving process from that of the intervention effects. This is because a crucial assumption in school consultation is that a good consultative process

leads to the development of a good intervention. This assumption is reflected in the literature in the fact that consultation research, to maintain internal validity, may hold interventions across participants constant. However, intervention quality is a complex, under-examined variable crucial to student outcomes. The current study has provided important contributions to the literature in that this research has provided preliminary validation evidence for a brief measure of intervention quality that might capture some of the complexity associated with it through investigation of the constructs, items, and reliability of the measure. Validation is not a singular event so this study is providing the groundwork for an instrument that can be reliably used to attend to intervention quality within consultation research. Further validation efforts are necessary to support its adoption.

References

Bahr, M. W., Leduc, J. D., Hild, M. A., Davis, S. E., Summers, J. K., & McNeal, B. (2017). Evidence for the expanding role of consultation in the practice of school psychologists. *Psychology in the Schools, 54*(6), 581-595.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37.

Bartlett, M. S. (1954). A further note on the multiplying factors for various chi-square approximations in factor analysis. *Journal of the Royal Statistical Society, Series B, 16*, 296-298.

Bellinger, S. A., Lee, S. W., Jamison, R., & Reese, M. (2016). Conjoint behavioral consultation: Community-school collaboration and behavioral outcomes using multiple baseline. *Journal of Educational and Psychological Consultation, 26*(2), 139-165.

Bergan, J. R. (1977). *Behavioral consultation*. Charles E. Merrill.

Bergan, J. R., & Kratochwill, T. R. (1990). *Behavioral consultation and therapy*. Plenum Press.

Bice-Urbach, B., & Kratochwill, T. R. (2016). Teleconsultation: The use of technology to improve evidence-based practice in rural communities. *Journal of School Psychology, 56*, 27-43.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*(149), 1-18.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores." *Journal of Educational Measurement, 50*(1), 74-83.

Breivik, E., & Olsson, U. H. (2001). Adding variables to improve fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 169-194). Scientific Software International.

Briesch, A. M., Chafouleas, S. M., Neugebauer, S. R., & Riley-Tillman, T. C. (2013). Assessing influences on intervention implementation: Revision of the Usage Rating Profile-Intervention. *Journal of School Psychology, 51*, 81-96.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*, 25-56.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press

Castillo, J. M., & Curtis, M. J. (2014). Best practices in systems-level change. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Systems-level services* (6th ed., pp. 11–28). National Association of School Psychologists.

Caplan, G. (1970). *The theory and practice of mental health consultation*. Basic Books.

Caplan, G., & Caplan, R. B. (1993). *Mental health consultation and collaboration*. Jossey Bass Publishers.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., & McCoach, D. B. (2009). Moving beyond assessment of treatment acceptability: An examination of the factor structure of the Usage Rating Profile - Intervention (URP-I). *School Psychology Quarterly, 24*, 36–47.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*(6), 560–575.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23-45.

Domenech Rodríguez, M. M., Baumann, A. A., & Schwartz, A. L. (2011). Cultural adaptation of an evidence based intervention: From theory to practice in a Latino/a community context. *American Journal of Community Psychology, 47*, 170-186.

Erchul, W. P., & Martens, B. K. (2010). *School consultation: Conceptual and empirical bases of practice* (3rd ed.). Springer Science + Business Media.

Erchul, W. P., & Sheridan, S. M. (2014). Overview: The state of scientific research in school consultation. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation* (2nd ed., pp. 3-17). Routledge/Taylor & Francis Group.

Erchul, W. P. (1987). A relational communication analysis of control in school consultation. *Professional School Psychology, 2*, 113-124.

Erchul, W. P., & Fischer, A. J. (2018). Consultation. In S. L. Grapin & J. H. Kranzler (Eds.), *School psychology: Professional issues and practices* (pp.181-195). Springer.

Erchul, W. P., Raven, B. H., & Wilson, K. E. (2004). The relationship between gender of consultant and social power perceptions within school consultation. *School Psychology Review, 33*, 582-590.

Erchul, W. P., & Ward, C. S. (2016). Problem-solving consultation. In S. R. Jimerson, M. K. Burns & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 73–86). Springer Science+Business Media.

Erchul, W. P., & Young, H. L. (2014). Best practices in school consultation. In A. Thomas & P. L. Harrison (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (6th ed., pp. 449–460). National Association of School Psychologists.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299.

Fischer, A. J., Dart, E. H., Radley, K. C., Richardson, D., Clark, R., & Wimberly, J. (2017) An evaluation of the effectiveness and acceptability of teleconsultation. *Journal of Educational and Psychological Consultation, 27*(4), 437-458.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299.

Forman, S. G., Fagley, N. S., Dreitlein Steiner, D., & Schnieder, K. (2009). Teaching evidence-based interventions: Perceptions of influences on use in professional practice in school psychology. *Training and Education in Professional Psychology, 3*(4), 226-232.

Frank, J. L., & Kratochwill, T. R. (2014). School-based problem-solving consultation: Plotting a new course for evidence-based research and practice in consultation. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation* (2nd ed., pp. 3-17). Routledge/Taylor & Francis Group.

Freer, P., & Watson, T. S. (1999). A comparison of teacher and parent acceptability ratings of behavior and conjoint behavioral consultation. *School Psychology Review, 28*(4), 672-684.

Garbacz, A. S., & McIntyre, L. L. (2015). Conjoint behavioral consultation for children with autism spectrum disorder. *School Psychology Quarterly, 31*(4), 450-466.

Garbacz, A. S., Watkins, N. D., Diaz, Y., Barnabas, E. R., Schwartz, B., & Eiraldi, R. (2017). Using conjoint behavioral consultation to implement evidence-based practices for students in low-income urban schools. *Preventing School Failure: Alternative Education for Children and Youth, 61*(3), 198-210.

Gersten, R., Baker, S., & Lloyd, J. W. (2008). Designing high quality research in special education: Group experimental design. *The Journal of Special Education, 34*(1), 2-18.

Gresham, F. M. (1982). *Handbook for behavioral consultation*. Iowa Dept. of Public Instruction.

Goh, A. E., & Bambara, L. M. (2012). Individualized positive behavior support in school settings: A meta-analysis. *Remedial and Special Education, 33*(5), 271-286.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265–275.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Ingram, K., Lewis-Palmer, T., & Sugai, G. (2005). Function-based intervention planning: Comparing the effectiveness of FBA function-based and non-function-based intervention plans. *Journal of Positive Behavior Interventions, 7*(4), 224–236.

Jenrich, R. L., & Sampson, P. F. (1966). Rotation for simple loading. *Psychometrika, 31*, 313-323.

Johns, R. (2005). One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion & Parties, 15*(2), 237-264.

Johnson, L. D., Fleury, V., Ford, A., Rudolph, B., & Young, K. (2018). Translating evidence-based practices to usable interventions for young children with autism. *Journal of Early Intervention, 40*(2), 158-176.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31-36.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457.

Kazdin, A. E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis, 13*, 259–273.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed). Guilford Press.

Kratochwill, T. R., Elliott, S. N. & Stoiber, K. C. (2002). Best practices in school-based problem-solving consultation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (4th ed., pp. 583–608). National Association of School Psychologists.

Krosnick, J. et al. (2002). The impact of 'no opinion' response options on data quality. *Public Opinion Quarterly, 66*, 371–403.

Lambert (2004). Consultee-centered consultation. An international perspective on goals, process, and theory. In N. M. Lambert, I. Hylander, & J. Sandoval (Eds.), *Consultee-centered consultation: Improving the quality of professional services in schools and community organizations* (pp. 3-19). Erlbaum.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949.

Lynn, M.R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(8), 382-5.

MacLeod, I. R., Jones, K. M., Somers, C. H., & Havey, J. M. (2001). An evaluation of the effectiveness of school-based behavioral consultation. *Journal of Educational and Psychological Consultation, 12*(3), 203-216.

Mannino, F. V., & Shore, M. F. (1975). The effects of consultation: A review of the literature. *American Journal of Community Psychology*, *3*, 1–21.

Martens, B. K., Witt, J. C., Elliott, S. N., & Darveaux, D. X. (1985). Teacher judgments concerning the acceptability of school-based interventions. *Professional Psychology: Research and Practice, 16*, 191–198.

Martens, B. K., DiGennaro-Reed, F. D., & Magnuson, J. D. (2014). Behavioral consultation: Contemporary research and emerging challenges. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation* (2nd ed.; pp. 210–247). Routledge.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23*, 311–318.

Medway, F. J. (1979). How effective is school consultation? A review of recent research. *Journal of School Psychology*, *17*(3), 809-818.

Mintzberg, H. (1983). *Power in and around organizations*. Prentice-Hall.

Molleman, G. R. M., Peters, L. W. H., Hosman, C. M. H., Kok, G. J., & Oosterveld, P. (2006). Project quality rating by experts and practitioners: Experience with Preffi 2.0 as a quality assessment instrument. *Health Education Research, 21*(2), 210-229.

Nadeem, E., Gleacher, A., & Beidas, R. S. (2013). Consultation as an implementation strategy for evidence-based practices across multiple contexts: Unpacking the black box. *Administration and Policy in Mental Health and Mental Health Services, 40*, 439-450.

National Association of School Psychologists (2022). *The professional standards*. https://www.nasponline.org/x55315.xml

Noell, G. H., & Gansle, K. A. (2014). Research examining the relationships between consultation procedures, treatment integrity, and outcomes. In W. P Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation* (2nd ed., pp. 386–408). Routledge.

Nowlis, S. M., Kahn, B. E., & Dhar, R. (2002). Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments. *Journal of Consumer Research, 29*(3), 319–334.

Ohmstede, T. J., & Yetter, G. (2015). Implementing conjoint behavioral consultation for African American children from a low SES urban setting. *Journal of Educational and Psychological Consultation, 25*(1), 18-44.

Paas, F., Tuovinen, J. F., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Powers, J. D., Bowen, N. J. & Bowen, G. L. (2010) Evidence-based programs in school Settings: barriers and recent advances. *Journal of Evidence-Based Social Work, 7*(4), 313-331.

Pua, D. J., Peyton, D. J., Brownell, M. T., Contesse, V. A., & Jones, N. D. (2021). Preservice observation in special education: A validation study. *Journal of Learning Disabilities, 54(*1), 6-19.

Reddy, L. A., Barboza-Whitehead, S., Files, T., & Reddy, L. A. (2000). Clinical focus of consultation outcome research with children and adolescents. *Special Services in the Schools, 16*(1), 1-22.

Revelle, W. & Rocklin, T. 1979, Very Simple Structure: An Alternative Procedure for Estimating the Optimal Number of Interpretable Factors. *Multivariate Behavioral Research, 14*, 403-414.

Raykov T., & Marcoulides G. A. (2011) *Introduction to Psychometric Theory*. Routledge, Taylor & Francis Group.

Rosqvist, J., Thomas, J. C., & Truax, P. (2011). Effectiveness versus efficacy studies. In J. C. Thomas & M. Hersen (Eds.), *Understanding research in clinical and counseling psychology* (pp. 319–354). Routledge.

Sandoval, J. (2014). Best practices in school-based mental health/consultee-centered consultation by school psychologists. In P. Harrison & A. Thomas (Eds.), *Best practices in school: Data-based and collaborative decision making* (pp. 493–507). National Association of School Psychologists.

Sanetti, L. M. H., & Gritter, K. L. (2010). Evidence based interventions: Resources and guidance for educators. In A. Canter, L. Paige, & S. Shaw (Eds.), *Helping Children at Home and School* (3rd ed.). National Association of School Psychologists.

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

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

Sanetti, L. M. H., Kratochwill, T. R., Collier-Meek, M. A., & Long, A. C. J. (2014). *PRIME manual: Planning realistic implementation and maintenance by educators.* Neag School of Education, University of Connecticut. Retrieved from www.implementationscience.uconn.edu

Sheridan, S. M. (1997). Conceptual and empirical bases of conjoint behavioral consultation. *School Psychology Quarterly, 12*, 119-133.

Sheridan, S. M., Clarke, B. L., & Ransom, K. A. (2014). The past, present, and future of conjoint behavioral consultation research. In W. P. Erchul & S. M. Sheridan (Eds.). *Handbook of research in school consultation* (2nd ed.; pp. 210–247). Routledge.

Sheridan, S. M., & Gutkin, T. B. (2000). The ecology of school psychology: Examining and changing our paradigm for the 21st Century. *School Psychology Review, 29*(4), 485–501.

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

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

Sheridan, S. M., & Steck, M. C. (1995). Acceptability of conjoint behavioral consultation: A national survey of school psychologists. *School Psychology Review,24*(4), 633-647.

Sullivan, A. L., & Long, L. (2010). Examining the changing landscape of school psychology practice: A survey of school-based practitioners regarding response to intervention. *Psychology in the School, 47*(10), 1059-1070.

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation-based assessment of clinical competence. *Advances in Health Science and Education, 23*, 323-338.

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.

Velicer, W. (1976) Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41-71). Kluwer Academic.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3,* 231-251.

Von Brock, M. B., & Elliott, S. N. (1987). Influence of treatment effectiveness information on the acceptability of classroom interventions. *Journal of School Psychology, 25,* 131-144.

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219-246.

Wu, H., & Leung, S. (2017). Can Likert scales be treated as interval scales? - A simulation study. *Journal of Social Service Research*, *43*(4), 527-532.

Zins, J. E., & Erchul, W. P. (2002). Best practices in school consultation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (4th ed., pp. 625-643). The National Association of School Psychologists.

Appendix A

Sample Content Validation Items on Qualtrics

Appendix B

Behavior Intervention Plan Sample

Student Background:

Name:	Jackson	**Gender:**	Male
Age:	8	**Race/Ethnicity:**	Black
School District geography:	Rural / Urban / **Suburban**	**Primary Language spoken in home:**	English
School District / Community SES:	**Affluent** / Middle / Low	**Secondary Language spoken in home:**	None
Student/Family SES:	Affluent / **Middle** / Low	**ELL / ESL enrolled:**	Yes / **No**
Student Disability Status:	Yes / **No**	**School / District resources availability:**	Low / Moderate / **High**
Overall Academic Achievement:	At grade level	**Staff/Student Ratio:**	1:23
Parent Involvement level:	Low / **Medium** / High	**Avg. Students per classroom:**	23
Parent Education Level:	No Diploma/GED GED Some College **Associates/Technical** Bachelor Masters PhD		

Formal FBA: Yes / **No**

Hypothesized Function:

X	**Escape/Avoid - Task**		Get/Obtain - Attention
	Escape/Avoid - Attention		Get/Obtain - Tangible
	Escape/Avoid - Task - Acquisition		
	Escape/Avoid - Task - Fluency		
	Escape/Avoid - Task - Generalization		

Jackson's Behavior Support Plan

The Problem

The purpose of the behavior:
Jackson avoids the demands of activities that he finds demanding (i.e., structured language-based activities, sharing objects, interactive play) by resisting or withdrawing. If pushed to participate, Jackson may react by throwing objects, screaming, or stating "shut up." During activities that are particularly demanding for Jackson, he may also show increased rigidity about favorite activities, objects, play routines, and conversations.

Things to Do All the Time

These strategies will assist Jackson in meeting the demands of difficult activities and social interactions.

Teach Play Skills – Support providers should enter into play activities and teach Jackson new play routines. Adults can provide support by scaffolding Jackson's interactions in play routines and centers. For example, an adult could invite Jackson into a center and then model or direct his play with peers.

Teach Social Interaction Skills – Adults should assist Jackson with turn-taking interactions by moving into play activities and mediating his social exchanges. For example, sit with Jackson and a peer. Tell the peer you want him to help you teach Jackson to play with ____. Give the toy to the peer. Cue Jackson to attend. "Look Jackson, Emily is pouring tea." Then cue Jackson to take a turn. "Jackson, you pour tea."

Teach Communication Repair Strategies – Adults should facilitate the use of communication repair strategies by Jackson. Currently, Jackson may mumble an answer if the adult fails to interpret his message or may exhibit problem behavior. Strategies that may be used include:

1. Interpreting his actions as if they are communicative (e.g., "Jackson, you're upset because Joey is in the name chair. Tell me, 'I want a name chair.' Jackson, I understand. You want a name chair. They are all gone. You can sit here or here.")

2. Asking for clarification when you don't understand what is said (e.g., "Tell me more") or asking that he demonstrates or use an object to show you (e.g., "Jackson, show me")

3. Repeat a portion of what he has said to acknowledge his message prior to asking for more clarification (e.g., "You are telling me the blocks are wrong. Tell me more.")

4. Create neutral opportunities for repairs by holding out for a repetition or modification of the language request (e.g., Jackson says "more please" at snack time. The teacher says "Jackson, more what please. Say I want." Then the teacher pauses for an additional response.)

Short Term Prevention Strategies

These strategies are used prior to situations that usually evoke resistance from Jackson

Choices – Choices should be given to Jackson throughout the day. Sometimes a concrete object or picture should be used to make sure that Jackson is making an informed choice and that he follows through with the choices he makes.

Personal Cuing – When Jackson is cued, those prompts need to be personal (i.e., directly given to him in a simple language) and understandable (paired with a gesture or object).

Simple Language – Jackson may have difficulty understanding complex language. It is especially difficult for him when he is upset or resistant to the proposed activity. On those occasions, use very simple language with Jackson when cueing him. Pair words with gestures (e.g., Pat chair and say "Sit. Sit in the chair.") or concrete objects.

Safety Signal – Jackson will be prepared for transitions that are going to be difficult by the use of a safety signal. Support providers will tell him "Jackson, in 5 minutes we will _____." Then cue him again at 3 minutes, and then transition will follow in 3 minutes by the support provider stating, "time for _____." Do not let him delay the transition. Follow through once the cue has been given.

Comfort Area – Jackson will be offered to move to a comfort area when he becomes frustrated or hurt. Visual symbols of emotions (e.g., frustrated, angry, sad, tired, sick) will be available and can be used by the adult and Jackson to help him label what he is feeling.

Positive Reinforcement – Jackson should receive statements about appropriate behavior frequently throughout the day in a natural fashion (e.g., "this is fun, I like playing with you").

Replacement Skills

Learn to negotiate difficult social situations – Social stories will be developed and used to help Jackson identify social cues, introduce new routines and rules, and teach him the social skills necessary for interactive play.

Learn to cope with negative emotions – Jackson will be assisted to identify the emotions that he is feeling. He may use picture symbols in the comfort area to discuss his emotions when he is upset. Choice option cards will be developed to assist Jackson in reacting to difficult situations without using problem behavior.

When the Problem Behaviors Happen

If Jackson has difficulty with moving into a new activity, use a language label. "This is hard for you." Then follow with the steps he needs to take. (e.g., "Sit in the chair. Get paper. Choose a marker.").

If Jackson becomes upset, encourage him to label what he is feeling and then use choice option cards to guide him in coping with the situation.

If Jackson becomes upset and needs a break, remind him that he can go to the comfort area. Guide him to select the visual that describes his emotion and facilitate his verbal expression of the emotion.

If Jackson yells “Shut up,” respond to him in a soft voice. State for him, “Jackson is feeling angry.” Then ask him to elaborate. “Tell me what happened” or “Tell me more.” Or, if Jackson seems unapproachable, prompt him to go to the comfort area.

Appendix C

Brief Measure of Intervention Quality (BMIQ)
High-Quality Intervention Components

Functionally Matched

1. The intervention clearly connects to a hypothesized function maintaining or creating the problem (i.e., seek/obtain, escape/avoid - acquisition, fluency, generalization).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

2. Overall, the intervention procedures appear to be appropriate for the presenting concern(s).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

3. The intervention is appropriate for the student's knowledge, skills, and abilities.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

4. The match between intervention mechanisms and desired outcomes is intuitive (i.e., the intervention appears likely to improve the targeted student functioning/behavior).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

5. The intervention is developmentally appropriate for the student/client.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

Procedurally Clear

6. Implementation procedures are clearly presented.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

7. Implementation procedures are thoroughly explained.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

8. The role of the interventionist is clearly defined.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

Evidence-Based

9. The intervention appears grounded in easily identified or recognized theoretical perspective (e.g., behaviorism, cognitive-behavioral, direct instruction)

Strongly disagree 0 1 2 3 4 5 Strongly Agree

10. The intervention appears grounded in a widely-accepted theoretical perspective (e.g., behaviorism, cognitive-behavioral, direct instruction).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

11. The intervention is clearly supported by a solid evidence-based (i.e., research).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

12. The intervention is consistent with best-practice recommendations for addressing the problem/difficulty of concern.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

Economical

13. Intervention implementation relies on an accessible and reasonable amount of human resources (i.e., staffing).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

14. Adequate resources (e.g., equipment, materials, notes) to support the implementation are available and accessible for the duration of the intervention

Strongly disagree 0 1 2 3 4 5 Strongly Agree

15. Intervention implementation relies on an accessible and reasonable amount of material resources (e.g., workbooks, flashcards, tangible rewards, technology).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

Acceptable

16. The intervention appears too complex/difficult to implement (e.g., numerous components or elements).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

17. Overall, the intervention appears to be one that an interventionist/teacher/parent/educator would be willing to use.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

18. The intervention appears implementable for most educators.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

19. The intervention is appropriate for the intended setting/context (e.g., school, home).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

<u>Culturally Sensitive/Adaptive/Appropriate</u>

20. The intervention is appropriate for the intended student/client (e.g., age, needs, disability, eligibility, background).

Strongly disagree 0 1 2 3 4 5 Strongly Agree

21. The intervention is appropriately sensitive/adaptive/responsive to the student/client's disability/ability status.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

22. The intervention is appropriately sensitive/adaptive/responsive to the student/client's language or communication skills and abilities.

Strongly disagree 0 1 2 3 4 5 Strongly Agree

23. The intervention is appropriately sensitive/adaptive/responsive to the student/client's racial/ethnic background.

Strongly disagree 0 1 2 3 4 5 Strongly Agree