2011

EVALUATION OF A COMMUNITY MENTAL HEALTH CENTER'S ASSESSMENT CLINIC: DEVELOPMENT OF A NOVEL ASSESSMENT EVALUATION TOOL

Cassidy C. Arnold
Virginia Commonwealth University

EVALUATION OF A COMMUNITY MENTAL HEALTH CENTER'S ASSESSMENT CLINIC: DEVELOPMENT OF A NOVEL ASSESSMENT EVALUATION TOOL

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University

By: CASSIDY C. ARNOLD
Bachelor of Arts; University of Colorado, Denver; May, 2006

Director: Michael A. Southam-Gerow, Ph.D.
Associate Professor
Department of Psychology and Pediatrics

Virginia Commonwealth University
Richmond, Virginia
December, 2011

Acknowledgements

I am deeply appreciative to the many people who have provided guidance and support throughout my time working on this project. Much of the academic guidance was provided by members of my thesis committee: Drs. Southam-Gerow, McLeod, and Geller. I would especially like to thank Dr. Southam-Gerow for his willingness to support my interests, read countless drafts, and hold many conceptual and pragmatic planning meetings.

I would like to thank my fellow lab members and graduate student peers for their logistical support and encouragement. On the logistical front, Shannon, Emily, and Lily graciously offered their time and helped with coding. I could not have completed this project without help from my many graduate student peers. At every turn and challenge, they provided valuable encouragement to push on and guidance for how to balance the many demands of graduate school.

I am grateful to my wonderful family—especially, Rick, Johanna, and Liliana; Teri, Dan, and Dara; Lloyd and Joan—for supporting my wandering academic and other pursuits that have led me to this point. Finally, many thanks and love to Lily--my cheerleader, right hand, and crashpad—for showing more tolerance, love, and encouragement than I could have ever hoped for. You are truly the best.

Table of Contents

Page

Acknowledgements ii

Tables of Contents. iii

List of Tables vii

Abstract ix

Introduction 1

Introduction to Assessment. 7

What is Assessment? 8

Clinical Decision Making. 11

What is Clinical Judgment? 12

Biases. 12

What is Actuarial Judgment? 13

Evaluating Psychological Instruments. 13

Measure Development 14

Readability. 14

Standardization. 15

Normative Data. 16

Reliability. 18

Test-Retest Reliability. 19

Internal Consistency. 20

Inter-Rater Reliability. 23

Alternate Form Reliability. 26

Validity. 27
What is a Construct?. 27
Content Validity. 29
Face Validity. 29
Construct Validity 30
Treatment Utility 34
Assessment Evaluation Systems. 35
'Technical Recommendations for Psychological Tests and Diagnostic Techniques'. 35
'Measures of Social Psychological Attitudes'. 37
'A Guide to Assessments that Work'. 38
Statement of Purpose. 39
Hypotheses 40
Method 42
Overview. 42
Participants. 42
Procedures. 43
Identification of the Sample. 43
Data Extraction and Coding. 44
Non-coded Variables. 44
Coded Variables. 51
Strength of Measure (SoM). 52
Informant and Method of Assessment. 58

Reason for Referral. 59

Referral and Diagnostic Domains Assessed. 61

Diagnostic Category. 62

Calculated Variables. 63

Coding and Data Entry Procedures. 63

Analytic Plan. 64

Results 65

Overview. 65

Preliminary Analyses. 66

Assessment Level Variables. 66

Strength of Measure (SoM). 66

Informant and Method of Assessment 69

Referral and Diagnostic Domains Assessed. 70

Diagnosis. 72

Primary Analyses. 73

Hypothesis 1: Assessment Instrument Adequacy 73

Hypothesis 2: Multiple Instruments Assessing

Reason for Referral 73

Hypothesis 3: Instrument Assessing Assigned Diagnosis 73

Hypothesis 4: Multiple Informants 73

Hypothesis 5: Multiple Methods 74

Post-hoc Tests 74

Post-hoc Analysis of Assessment Instrument Adequacy 75

Post-hoc Analysis of Multiple Instruments

Assessing Reason for Referral 75

Post-hoc Analysis of Instruments

Assessing Primary Diagnosis Domain. 79

Discussion 81

Strength of Measure and Evaluation of Assessment Instruments. 82

Comparing the Result of the SoM to Other Assessment Reviews. 86

Strengths and Weaknesses of the Reports. 89

Assessing Reason for Referral 89

Assessing Primary Diagnosis 90

Informants and Method of Assessment 90

Limitations 92

Future Directions. 96

Conclusion. 98

List of References 100

Appendices 107

A Strength of Measure Criteria. 100

B Strength of Measure Coding Sheets. 113

C Reason for Referral Coding Sheets. 129

Vita. 132

List of Tables

		Page
Table 1.	Youth Demographics.	44
Table 2.	Non-coded Variables.	45
Table 3.	Assessment Component Data.	47
Table 4.	Primary Axis I Diagnosis Category.	51
Table 5.	Coded Variables.	51
Table 6.	Literature Review.	54
Table 7.	Informant.	58
Table 8.	Method of Assessment.	59
Table 9.	Reason for Referral.	60
Table 10.	Referral Domains Assessed.	61
Table 11.	Diagnosis Category and Diagnosis Domains Assessed.	62
Table 12.	Calculated Variables.	63
Table 13.	Assessment Level Variables.	67
Table 14.	SoM Results.	71
Table 15.	Reason for Referral and Diagnostic Domains Assessed.	72
Table 16.	Reason for Referral Analyses.	77
Table 17.	Diagnosis Category Analysis.	81
Table 18.	Strength of Measure Variable: Readability.	108

Table 19. Strength of Measure Variable: Standardization. 108

Table 20. Strength of Measure Variable: Normative Data. 109

Table 21. Strength of Measure Variable: Internal Consistency. 109

Table 22. Strength of Measure Variable: Inter-rater Reliability. 110

Table 23. Strength of Measure Variable: Test-retest Reliability. 110

Table 24. Strength of Measure Variable:
Construct/Representative Validity. 111

Table 25. Strength of Measure Variable:
Criterion-related/Elaborative Validity. 112

Table 26. Strength of Measure Variable: Treatment Utility. 112

Abstract

EVALUATION OF A COMMUNITY MENTAL HEALTH CENTER'S ASSESSMENT CLINIC: DEVELOPMENT OF A NOVEL ASSESSMENT EVALUATION TOOL

By Cassidy C. Arnold, BA.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2011

Major Director: Michael A. Southam-Gerow, Ph.D.
Associate Professor
Clinical Psychology

High quality assessment services are the linchpin connecting youth with mental health problems to diagnosis-specific, evidence-based treatments. The effort to improve assessment services is in its early days and faces a number of substantial challenges. This study was an initial effort to address these challenges through the development of a standardized, multi-dimensional *Assessment Instrument* evaluation tool—the Strength of Measure (*SoM*)—based on operationally defined criteria supported by decades of psychometric research. The *SoM* and other criteria addressing assessment practices were piloted with data from 32 consecutive assessment reports from a community mental health center's Assessment Clinic. Results indicate that none of the *Assessment Instruments* used by the Assessment Clinic met the "*Adequate*" level of support on

each of the *SoM* dimension. Additional results address *Reason for Referral*, *Primary Axis I Diagnosis*, *Informants*, and *Method of Assessment*. Implications and directions for future research are discussed.

Evaluation of a Community Mental Health Center's Assessment Clinic: Development of a Novel Assessment Evaluation Tool

One in five children and adolescents have some type of mental health problems in any given year and half of all mental health disorders affecting adults have an onset before the age of 14 (Mash & Hunsley, 2005; Mash, 2007). Effective care for youth and adults with mental health disorders is a financial imperative. The cost of mental health disorders in the United States includes approximately $12 billion dollars a year for children alone (Mash, 2007). Additionally, there are long-term consequences that result from mental health problems that are difficult to enumerate; some consequences include academic failure, underperformance, and loss of productivity (Mash, 2007).

As of the 1960s, there were relatively few treatment paradigms (psychoanalytic, client-centered, and behavioral; Kiesler, 1966; Hayes, Nelson, & Jarrett, 1987). Psychotherapists trained in these practices, treated most clients using the same treatment model with little variation between clients despite differences in presenting problems (Nelson-Gray, p. 521, 2003; Hayes, Nelson, & Jarrett, 1987). As more and more problem-specific treatments have been developed (Hayes, Nelson, & Jarrett, 1987; Weisz & Addis, 2006) the myth that one treatment model works equally well for all patients and problems has faded (Kiesler, 1966; Chambless & Hollon, 1998). That is, the belief that one treatment paradigm is the best treatment for all patients regardless of their presenting problem and personal characteristics has transitioned toward emphasizing treatment selection in light of the patient's characteristics: most notable, their diagnosis (Chambless & Hollon, 1998). This is reflected in the APA policy statement on Evidence Based Practice in Psychology which emphasizes combining research, clinical

expertise, and patient characteristics (APA, 2005 a). However, despite advances in diagnostic and problem behavior specific treatments, research on assessment practices to support evidence-based treatment selection has not kept pace with treatment development (Achenbach, 2005; Mash & Hunsley, 2005).

Over the past couple of decades, substantial work in research and clinical settings has improved the quality of treatment for youth with a variety of mental health problems (Mash, 2007). As the field of psychology develops more and more treatment approaches with demonstrated efficacy (e.g., cognitive behavioral therapy for anxiety and Multisystemic Therapy for externalizing behaviors) designed for clinical populations with specific characteristics (e.g., a particular disorder, problem behavior, or skill deficient), the need for high quality assessment practices becomes more important (Hayes, Nelson, & Jarrett, 1987; Nelson-Gray, 2003). That is, as more treatment options are developed and the more that these treatment options are specialized to treat a particular problem or disorder (Weisz & Addis, 2006), the importance of accurately differentiating between mental health disorders for the sake of treatment selection becomes increasingly necessary (Mash, 2007; Barry & Pickard, 2008). Despite this need, assessment practices have received far less research attention than treatment options (Hayes, Nelson, & Jarrett, 1987; Mash & Hunsley, 2005; Achenbach, 2005).

If clinicians were not able to identify the relevant characteristics of the patient then they will not be able to take full advantage of the advances in treatment for specific populations of patients (Weisz & Addis, 2006). It has been recognized for some time that individuals entering therapy are not a uniform group and that patient characteristics have real implications for response to treatment (e.g., Kiesler, 1966; Barkley, 1997). For example, with respect to demographic characteristics, younger children with defiant behavior respond better to parent

management training than do adolescents (Barkley, 1997). Additionally and more germane to this paper, exposure-based therapy alone is appropriate for individuals with social anxiety but without a social skills deficit whereas individuals with social anxiety and social skills deficit do not benefit as much from exposure alone but require social skills training in order to see a reduction in symptoms (Weisz & Addis, 2006). In this case, knowing whether an individuals' social anxiety is accompanied by a social skills deficit is important when planning their treatment. This distinction must be made via some form of assessment. Therefore, advancing the quality of assessment instruments available to clinicians, and clearly articulating the purpose and method of use for each assessment instrument, is a necessary link between treatment development in research settings and improving the outcomes for individuals with mental health problems.

The criteria published by Division 12 for establishing evidence-based treatments requires that participants in the research trials belong to an identifiable group (e.g., a specific diagnosis) but does not specify what assessment procedures and instruments should be used to establish such group membership (APA Division 12, 1995; Chambless & Ollendick, 2001). Chambless and Hollon's (1998) pivotal article on treatment evaluation and the development of a set of criteria for establishing evidence-based treatments (APA, 1993), states that treatments should not be broadly evaluated for their level of empirical support but, rather, should be evaluated for their empirical support in the context of a particular disorder. Inherent in this statement is that treatments are not uniformly effective for all patients or disorders and, therefore, some form of assessment and problem identification is necessary. As such, it is incumbent on the field of psychology to identify which characteristics, including but not limited to mental health disorders

and problems, are relevant when selecting a treatment and how these characteristics should be assessed.

As mentioned earlier, the emphasis of the importance of patient characteristics on response to treatment articulated by Kiesler (1966), Chambless and Hollon (1998), and others is reflected in the APA's policy statement on evidence-based practice in psychology (APA, 2005a; APA, 2005b). In that statement, the APA defines evidence-based practice in psychology as the convergence of research, clinical experience, and patient characteristics. Furthermore, just as there is a need for developmentally appropriate treatments for children based on their age and developmental level, there is also a need for developmentally appropriate assessment strategies to support treatment selection (Holmbeck, Greenley, & Franks, 2003). A thorough understanding of the normative developmental stages and milestones is critical to understanding, assessing, and recommending interventions for youth referred for psychological assessments (Barry & Pickard, 2008, pp. 77-78). For example, meta-analysis has shown that the effect size of cognitive behavioral therapy, which requires a degree of abstract thinking, is twice as high among adolescents as in children (Holmbeck, Greenley, & Franks, 2003). An understanding of development can aid a clinician's understanding of how an adaptive and normative behavior at one age can become disordered later in life (Rutter & Sroufe, 2000).

The literature on case conceptualization articulates the role that assessment plays in the treatment process, from a slightly different perspective. Accurate and complete assessment allows the clinician to develop a case formulation that understands the relationships between variables (e.g., problems, antecedent conditions, behaviors, thoughts, emotions, and consequences) relevant to the patient, thereby facilitating treatment planning to specifically address the underlying cause of psychopathology (Freeman & Miller, 2002). Persons (1989)

advocates for collecting data from patient report and assessment measures. A case conceptualization, based on strong assessment data (supported by research on psychometrics and utility; discussed in greater detail later in this document in the "Evaluation of Psychological Instruments" section), guides the therapist's treatment planning decisions including identification of therapeutic priorities and choice of treatment modality and intervention strategies as well as ongoing treatment decisions (Pearsons, 1989; Freeman & Miller, 1992). In other words, based on this perspective, the therapist who most fully understands the patient's presenting problem and how those problems are interacting with the patient's internal world (e.g., thoughts, emotions, physiology, and brain activity) and their environment is best able to select a treatment that will effectively treat the patient.

Sommers-Flannagan (2009) states that "assessment should inform treatment planning" (p. 295). Understanding the patient's problems and goals based on a systematic assessment approach helps to ensure that the foci of treatment and goals are comprehensive rather than based on first impressions or errors in judgment (see the section on Clinical decision making and Biases later in this document). Ensuring that patient problems that are relevant to treatment are identified and defined enables the clinician to develop intervention strategies that specifically target each problem. Research on suicidal adolescents suggest that directly focusing on the problem behavior (i.e., the suicidal thoughts and actions) leads to a greater reduction in suicidal acts than addressing the emotional state (e.g., depression or anxiety) believed to be the underlying and causing the problem behavior (Miller, Rathus, & Linehan, 2007). The case-conceptualization and diagnostic based treatment planning approaches both are predicated on the notion that assessment leads to better treatment outcomes and is the central question addressed by the research on the treatment utility of psychological assessment. Treatment utility of

psychological assessment addresses the benefits of psychological assessment with regard to clinical outcomes (Hayes, Nelson, & Jarrett, 1987; discussed in further detail later in this document).

Within the field of clinical child and adolescent psychology, there is currently a wide variety of assessment practices and measures designed to achieve a number of different purposes (Mash & Hunsley, 2005; Groth-Marnat, 2003). The present paper will focus on assessment practices that aid in diagnostic decisions and making treatment recommendations. Currently, there is no agreed upon method for evaluating assessment practices. Researchers advocating for the use of evidence-base assessment practices have proposed a number of different dimensions on which to evaluate assessment instruments including various forms of reliability, validity, standardization, and normative data (Mash & Hunsley, 2005).

This study evaluates the practices used at a community mental health clinic's (CMHC) Assessment Clinic. The first stage of the study addressed the development and application of a novel assessment evaluation tool, the *Strength of Measure* (*SoM*) that provides data on the extent to which each *Assessment Instrument* used by the Assessment Clinic conformed to the tenants of evidence-based assessment practices. Specifically, the *Assessment Instruments* used by the Assessment Clinic will be evaluated on nine dimensions: (a) *Readability*, (b) *Standardization*, (c) *Normative Data*, (d) *Internal Consistency*, (e) *Inter-rater Reliability*, (f) *Test-retest Reliability*, (g) *Construct/representative Validity*, (h) *Criterion-related/elaborative Validity*, and (i) *Treatment Utility*. The *SoM* will be reported for the *Assessment Instruments* most frequently used at the Assessment Clinic.

The following section will examine the body of research pertaining to the field of evidence-based assessment with an aim of defining dimensions that other researchers (e.g., Mash

& Hunsley, 2008; Mash & Hunsley, 2005; Robinson, Shaver, & Wrightsman, 1991, American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) have determined are needed to evaluate assessment efforts. First, a review of the assessment literature will provide a background on the purpose and methods of psychological assessment to identify the scope and role of assessment in clinical psychology. Second, clinical versus actuarial decision making will be reviewed to identify if and where evidence-based assessment procedures are needed to supplement clinical judgment. Third, the research on psychometrics and assessment evaluation systems will be reviewed in order to provide a background for the use of each element to be included in the evaluation tool developed for the present study. Following this background information, the specifics of the current study are described.

Introduction to Assessment

Clinical assessment is "the process by which clinicians gain understanding of the patient necessary for making informed decisions" (Korchin, 1976, in Hayes, Nelson, & Jarrett, 1987). In other words, psychological assessment is the process of gathering information about a patient's psychological and social problems, limitations, and capacities and his or her circumstances for the purpose of making decisions about how to best help the patient (Kendall & Norton-Ford, 1982). Clinicians must select means of collecting information; gather a large amount of information, sometimes from different sources; integrate information that is sometimes contradictory; interpret and summarize the information; and communicate the information so that effective interventions can be implemented (Kendall & Norton-Ford, 1982; Groth-Marnat, 2003). A psychological assessment goes beyond administering tests and listing traits or ability levels. A central goal of psychological assessment is to understand an

individual's strengths and weaknesses within the context their environment for the sake of developing and implementing a treatment plan to help them solve a problem or improve their functioning (Groth-Marnat, 2003). Because not every patient enters therapy for the same reason, with the same background, and with the same characteristics (e.g., level of cognitive development, trauma history, and social skills), assessment is an essential prerequisite to therapy (Kendall & Norton-Ford, 1982). Without high quality assessment, mental health services will too frequently target the wrong individuals, apply the wrong intervention, fail to modify treatment when needed, and, in the end, achieve sub-optimal outcomes.

What is Assessment?

To ensure a common understanding of the purposes and breadth of assessment, this section will identify some of the main uses of assessments, methods of assessment, and the strengths and weaknesses of various assessment method.

Psychological assessment may be used for a wide range of purposes. Some of these include: (a) screening (identifying which individuals require additional psychological assessment to determine if they have psychological problems that warrant treatment), (b) diagnosis and/or problem identification (categorizing individuals based on psychopathology and better understanding the individuals particular symptom constellation), (c) selecting or developing a treatment plan, (d) prediction of symptom course and/or treatment outcome, (e) ongoing treatment monitoring (evaluating the progress of a treatment by measuring changes in symptoms, functioning, and distress), and (f) evaluating treatment outcomes (determining if the treatment achieved its intended goals; Mash & Hunsley, 2005; Groth-Marnat, 2003). It is important to consider the assessment instrument's purpose, as an instrument designed for one purpose may

not be appropriate for other purposes (e.g., screening tools may not be as useful in developing a treatment plan).

Similarly, there are a wide variety of methods that can be used to conduct assessments. Commonly used assessment methods include clinical interviews with varying degrees of structure; individually administered self-report questionnaires; questionnaires completed by an individual familiar with the patient's current and past behaviors, moods, and thoughts; behavioral observation; self-monitoring; projective assessments; and review of past documentation of the patient's performance and medical, psychological, and academic evaluations (Groth-Marnat, 2003).

Each assessment method has strengths and weaknesses. These strengths and weaknesses make each type of assessment method well suited for some purposes but ill suited for others. For example, *self-report* and *informant-report* instruments, particularly brief instruments, lend themselves to screening individuals or assessing need for services because can be administered and scored by minimally trained individuals or on a computer, are often inexpensive, and can be administered to a large number of people. However, instruments designed for screening purposes tend to be highly sensitive. That is, since the task of screening is to ensure that as many people as possible who need services are identified, the assessments used for this purpose aim to identify a very high percentage of individual who truly need services and accept a high rate of false identification of individuals who do not need services (Groth-Marnat, 2003). Thus, instruments designed for screening purposes are ill suited for diagnostic assessment.

Self-report or *informant-report* instruments can also be used for diagnostic purposes, case conceptualization, treatment planning, treatment monitoring, or outcome measures. When used for these purposes, the instruments need to strike a balance between high sensitivity, described

above, and high specificity. Specificity refers to the rate at which individuals are accurately categorized who do not have a particular disorder or trait (Groth-Marnat, 2003). Whereas the purpose of screening instruments are to correctly identify as many individuals as possible who are in need of services, accepting false positives, diagnostic instruments are meant to only diagnose those individuals who truly have the disorder of interest. Self-report or informant-report instruments for diagnostic purposes tend to be more costly to administer and score than screening measures because of their longer length and higher level of expertise required to administer, score, and interpret.

Clinical interviews (e.g., Anxiety Disorder Interview Schedule, Kiddie-Schedule for Affective Disorder and Schizophrenia, and Diagnostic Interview Schedule for Children) are also used for diagnostic, case conceptualization, treatment planning, treatment monitoring, or outcome measures. Clinical interviews generally require a high level of clinical training and are time consuming. As a result, they are expensive. On the other hand, a well conducted clinical interview can provide rich information.

Self-monitoring and *behavioral observation* can provide valuable information on the relationship between different variables (e.g., thoughts, feelings, behaviors, internal and external stimuli, interpersonal relationship, physiology, and more) in the patient's life and can aid in treatment decisions. The strength of these measures is the detail that they can capture and their ability to be individualized for each patient. However, they each have serious limitations. Specifically, self-monitoring requires that the patient be aware of their behaviors, thoughts, and feelings, the circumstances in which these events occur, their consequences, and are able to record this information accurately. Therefore, the reliable of self-monitoring is limited by the reliability of the patient to be able to observe and report these events. Furthermore, the

interpretation of this information is not always straight forward. Behavioral observation can be an attempt to get around these challenges, assuming that the mental health worker or other reporter is able to reliably observe and record the relevant behavioral data, but is limited by high cost due to the time consuming nature of behavioral observation.

Broadly speaking, *projective assessment* instruments, such as the Rorschach and the Thematic Apperception Test, require the patient to view and respond to ambiguous stimuli. How an individual responds to these stimuli is, in theory, a reflection of their motivations, cognitive tendencies and structure, and their personality (Groth-Marnat, 2003). Unlike other assessment instruments, projective instruments do not depend on an individual's ability to reflect on their own thoughts, feelings, and behaviors and therefore is not limited by this variable. On the other hand, projective instruments are limited by the high degree of training required to administer, score, and interpret. Additionally, the reliability and validity (defined later in this document) have long been a concern for projective instruments such as the Rorschach (Groth-Marnat, 2003).

While recognizing that there are many different types of assessment that serve different purposes, the present study focuses on the assessment practices that guide diagnosis and treatment selection and the theory and evidence guiding these assessment practices. To arrive at a diagnosis, the clinician must integrate a large amount of information that has been collected through any number of the methods described above plus their unstructured interactions with the patient. There are no specific, empirically-based guidelines for how to integrate this information and diagnostic decision making and treatment planning can require that the clinician make difficult decisions. Because of the need for clinicians to integrate diverse data using their own judgment, the next section focuses on the literature related to clinical decision making.

Clinical decision making. This section will briefly discuss some of the rationale for highly structured assessment practices and instruments based on normative data. This will include a discussion of biases, clinical judgment, actuarial judgment, and how the two types of judgment should be integrated.

What is clinical judgment? One broad form of decision making is based on clinical judgment. Clinical judgments are decisions made by clinicians based on intuition and unsystematically weighted observations from various sources ranging from test and interview data, past records, and behavioral observations (Dawes, Faust, & Meehl, 1989; Groth-Marnat, 2003). Decisions made based on clinical judgment can but do not necessarily include data from formal assessment measures. Clinical judgments are susceptible to the biases discussed later in this document because they are based on idiosyncratic methods of aggregating, weighting, and applying meaning to various pieces of information. Biases on the part of the clinician and the shortcomings in clinical judgment represent real problems when it comes to providing the best possible service to individuals in need of psychological intervention. It should be clearly noted that, despite the similarity in terms, clinical decision making is not synonymous with actions by clinicians or even decisions made by clinicians since clinicians can make decision based on a clinical or actuarial model (Dawes, Faust, & Meehl, 1989).

***Biases*.** There is a substantial research in the field of clinical and cognitive psychology on decision making. One focus of this work is on the accuracy of highly trained professionals (clinicians) with varying degrees of experience, education, and clinically relevant data. This research has found clinicians are unable to reliably make many clinically relevant decisions. For example, clinicians tend to: (a) weight more heavily information gathered early in the data collection process (primacy effect), (b) over attend to data that confirms their diagnostic hunch

(confirmatory bias) and ignore disconfirming data, (c) aggregate data in idiosyncratic ways, and (d) base their decisions on the data that is available to them while ignoring the gaps in their knowledge (availability heuristic; Angold & Fisher, 1999; Groth-Marnat, 2003). Additionally, clinicians have a tendency to see correlations between variables even when there is not a correlation and see variables as independent when they are related, thereby missing real correlations (Angold & Fisher, 1999).

What is actuarial judgment? Actuarial or statistical judgment is decision making that is based on data and formulaic methods (Meehl, 1956). These formulaic methods include the use of cutoff scores and regression equations (Groth-Marnat, 2003). The difference between clinical judgment that uses measurement data and actuarial judgment is that clinical judgment is based on unsystematic weighing of data whereas actuarial judgment is based on systematic decision making rules. Given the exact some data, decisions based on clinical judgment do not necessarily result in a consistent conclusion whereas actuarial based decision making will result in the same conclusion (right or wrong) each time. Research over the past half century has generally supported Meehl's stance that actuarial judgments are superior to clinical judgments (Groth-Marnat, 2003).

Evaluating Psychological Instruments

Groth-Marnat (2003) advocates for a basic understanding by clinicians of assessment instruments prior to their use in clinical settings. Specifically, clinicians should understand (1) the domains the instrument measures; (2) the demands that the instrument places on the informant in terms of time and cognitive abilities; (3) how to administer and score the instrument; (3) the normative base on which the instrument has been evaluated; (4) the reliability of the measure; (5) the validity of the measure; and (6) the utility of the measure (e.g., Groth-

Marnat, 2003). However, given the vast number of assessment instruments and lack of uniformity in presentation of the information provided about psychological instruments, and time pressures under which clinicians operate, staying sufficiently abreast of the literature in order to select the instruments with the greatest level of empirical support is a daunting task (Mash & Hunsley, 2005). Therefore, a well organized, scientifically rigorous, and easily disseminated method for evaluating psychological measures will be beneficial to clinicians and, ultimately, their patients.

The following section will review some of the criteria by which psychological instruments can be evaluated and efforts to develop a comprehensive system for evaluating psychological instruments. Criteria for evaluating psychological instruments can be roughly organized into the following categories: (a) measure development, (b) normative data, (c) reliability, (d) validity, and (e) utility.

Measure development. Measure development includes two aspects measured by the *SoM*: *Readability* and *Standardization*. Both are unique in that they are established by the measure developers prior to any individuals being administered the *Assessment Instrument*. For that reason, it can be said that *Readability* and *Standardization* are fully within the control of the measure developers.

Readability. *Readability* refers to the reading level required to read and understand text. The readability of self-report measures has received recent attention as clinicians and service agencies are increasingly being asked to justify their fees (McHugh & Behar, 2009). Additionally, readability is a particular issue when working with individuals from populations with lower literacy rates such as the elderly, particular ethnic groups, and those with mental illness (McHugh & Behar, 2009). *Readability* can be measured in a number of different ways

but generally takes into account the number of syllables per word, the length of sentences, and the number of words used that are not found on basic word lists (McHugh & Behar, 2009). Based on findings that suggest that individuals with mental illness and minorities have lower than average reading ability, it is possible that the use of measures with too high a reading level are partially to blame for the disparity in mental health outcomes (McHugh & Behar, 2009). Furthermore, readability is important because of its impact on other metrics. That is, if an individual cannot read and comprehend the measure a clinician has asked them to complete, then they cannot reasonably be expected to provide reliable and valid data.

Standardization. *Standardization* refers to the structured and consistent procedure for administering, scoring, and interpreting a psychological assessment instrument (Anastasi, 1988). Ideally, the score on a psychological assessment would be purely the result of the underlying construct being measured. However, in reality, other variables affect the scores on the instrument. These variables are collectively referred to as error (DeVellis, 2003). The way in which the instrument is administered contributes to the error in a score. By standardizing the way in which a measure is administered, error is not eliminated but the amount of error resulting from variation in administration procedures is held reasonable constant and to a minimum. Therefore, the error associated with administration method can be accounted for in the scoring of the instrument. Additionally, standardization allows for the comparison of test scores across different administrations of the instrument with the same individual and between individuals (Anastasi, 1988). Aspects of administration that should be held constant include sequence of items, format of items (e.g., oral versus written), format of response (e.g., oral versus written, open ended response versus close ended response), and context of administration (e.g., group versus individual administration).

In addition to the method for administering the instrument, the method of converting an individual's response into a score must also be standardized for the same reasons. Sometimes scores on individual items are simply added together to arrive at a single score for the entire measure. Other times, scores from different items are weighted differently or added to together in various combinations in order to form multiple scores within a single instrument. By standardizing the method by which raw data is combined, scores on a measure can be compared across time within a single individual or between individuals or groups. By doing so, clinicians are able to assign meaning to a particular score beyond item level analysis (see the section on normative data later in this document). By comparing scores calculated in a standardized way from an individual with scores from individuals for which certain information is known (e.g., diagnosis, distress level, depression) clinicians are able to interpret what a particular score means.

Normative data. A measure is said to have normative data when the measure has been administered to a known group and, at a minimum, the mean and standard deviation of the scores for this group is provided. The accurate interpretation of scores is dependent on the development of normative data. Normative data is important whenever the responses to an assessment instrument are going to be used to make statements about the individual compared to other individuals. Any statement that directly makes a comparative statement, such as statements about the individual's percentile rank ("Sally's score is above the 75th percentile compared to other youth her age") or even what is considered normal ("Sally's difficulty concentrating is typical of individuals who have recently experienced a traumatic event") should be supported by normative data. Additionally, statements that imply a comparison (Sally's responses indicate that she has a strong imagination") should also be supported by normative data.

The quality of normative data comes down to three important issues. First, normative data should be developed on a randomly sampled group that is representative of the group or individual for which the assessment instrument is going to be used (Groth-Marnat, 2003). Normative data that are developed based on samples of convenience run the risk of being skewed systematically, thus adding additional systematic error into the interpretation of scorers. Second, norms should be based on a sufficiently large sample (Groth-Marnat, 2003). Bridges and Holler (2007) determined that research establishing normative data should use at least 50 participants per cell. Norms based on samples that are too small, even if random sampling was used, may not be representative of the population and lead to over or under pathologizing the results of psychological assessments (Bridges & Holler, 2007). Specifically, when too few individuals are included in the normative set, the standard deviation of the scores will be larger than if more individuals were included. This results in reduced confidence in the meaning of scores that deviate from the mean. Third, normative data should also be provided for subsamples of the population for which scorers may differ systematically from the scores of the population as a whole (Groth-Marnat, 2003). Frequently subsamples of the population are divided by gender, age, ethnicity, educational attainment, mental health status, and socioeconomic status (Groth-Marnat, 2003). For assessments of youth, these groups can include parental characteristics such as parental educational attainment or family income. Additionally, the normative data for assessments of youth should include narrowly defined age categories (e.g., 1-3 months) due to the rapid developmental changes in this stage of life. For example, in order for normative data for an assessment of aggressive or impulsive behavior to be useful, toddlers and adolescents should not be combined into the same group. This would have the result of over pathologizing a toddler's behavior and not catching problematic adolescent behavior.

Reliability. DeVellis (2003, p. 27) describes reliability as "the proportion of variance attributed to the true score of a latent variable". In other words, reliability refers to the degree to which an instrument consistently assesses the characteristic of interest versus other characteristics and error (Kazdin, 2003; Groth-Marnat, 2003). Reliability is typically measured by looking at test-retest or temporal stability of the instrument (correlation between scores on the same instrument at two different times), internal consistency (Cronbach's alpha, split-half, or parallel forms), and inter-rater reliability (raw agreement, Cohen's Kappa, Interclass Correlations, and others; Kendall & Norton-Ford, 1982).

Central to the issue of reliability is the concepts of true score, error variance or error, and error of measurement. In classical test theory, the true score refers to the score on an instrument of a given construct if the construct of interest was the only influence on the score (Sattler, 2008). A central assumption of classical test theory, item response theory, and other theories of measurement is that error, defined as all other influences on an instrument's score, is random and has a mean of zero across an infinite number of administrations to the same individual or across individuals (DeVellis, 2003; Anastasi, 1988). Therefore, the actual score that an individual receives on an item or instrument is equal to the true score plus the error variance (DeVellis, 2003). The relationship between the true score on an item or instrument and the error variance of that item or instrument is referred to as the error or measure with lower errors of measure corresponding to higher reliability (DeVellis, 2003).

Inherent in every psychological instrument is a certain degree of error due to any number of possible sources including inattention; miss-worded, misread and misunderstood items; variations in characteristics of the patient that impact scores on an instrument but are not the construct of interest; and unaccounted for noise that is inherent in measurement of every type. If

the errors are too great, then the score on an instrument will be overly influenced by variables other than variable of interest, making the instrument of little use. Such an instrument would be said to have unacceptably low reliability.

Groth-Marnat (2003) summarize the types of reliability as measuring reliability from time to time (test-retest reliability), item to item (internal consistency), and rater to rater (inter-rater reliability). Also discussed is form to form reliability (alternate forms reliability), but alternate forms reliability is not included in the final assessment evaluation criteria because the purpose of this study is to assess the quality of assessments conducted for diagnostic and treatment planning purposes and alternate forms of an instrument is less applicable in that context.

Test-retest reliability. *Test-retest reliability* or *temporal stability* refers to the degree to which scores on a measure at one point in time are similar to scores on the same measure at a second point in time (Kendall & Norton-Ford, 1982). *Test-retest reliability* is reported as a correlation of the two scores with higher scores indicating higher stability over time (Sattler, 2008). Changes in scores over time on the same measure are due to one or more of three reasons. First, real changes in the individual can affect the score. These changes may be in the construct of interest or another, unrelated constructs such as reading ability. Second, changes in the assessment taker's responses resulting from repeated administrations of the same test. This is frequently referred to as a practice effect. The third reason why scores may change over time is random variance or error in the individual's responses. While it is easy to recognize that random error is not a positive characteristic for an assessment measure, temporal stability is not a uniformly desirable assessment characteristic. For example, assessments that measure constructs that are known to change over time would be expected to have lower *test-retest reliability* than

assessments that measure stable. High *test-retest reliability* is desirable in measures of constructs that are presumed to be stable but would be a problem for measures that are used to assess constructs that are presumed to be variable or amenable to treatment. Instruments used to assess treatment effects with very high temporal stability could inadvertently mask treatment effects. When evaluating an instrument's *test-retest reliability*, it is also important to consider the length of time between the first and second administration of the instrument since it is more likely that an underlying construct changes over the course years versus days. A major limitation of assessing reliability via temporal stability is the difficulty explaining why a score changes over time. Again, a low correlation could be an accurate reflection of changes in the construct or measurement error.

Internal consistency. *Internal consistency* is another form of reliability. *Internal consistency* refers to the homogeneity of items within a scale (DeVellis, 2003). This criterion can be applied to a single subscale within an instrument or the instrument as a whole if the instrument measures only one dimension. It is desirable for subscale to have high internal consistency because that indicates that the items comprising the subscale are measuring a single underlying construct (DeVellis, 2003). When a scale is measuring a single construct, interpretation of scores from that scale are more straightforward. When scale has low internal consistency, the resulting score on that scale could be an indication that two or more underlying constructs are being measured. The result is that the interpretation of scores on a scale with low internal consistency is difficult because there are multiple explanations and score. It should be noted that tests of internal consistency, like other tests of reliability, indicate to what degree the items within an instrument reflect a single, underlying construct but does not test whether that

construct is the one in which the developers of the instrument intended to measure (DeVellis, 2003). That is a question of validity and is discussed later in this document.

There are a number of different ways to calculate the internal consistency (e.g., Cronbach's alpha, Kuder-Richardson, and split-half reliability) of an instrument but each method is based on the same fundamental premise that was described above. Split-half reliability is the oldest (Streiner, 2003) and perhaps the most conceptually straight-forward measure of internal consistency. For instruments that consist of items that measure only a single construct, the two halves of the instrument will measure the same construct and will therefore be highly correlated with one another. Split-half reliability is derived by administering an instrument one time, dividing the item pool into two, and correlating the scores (Groth-Marnat, 2003)[1]. The items making up the instrument can be divided randomly, by half (first half and last half), or by odd and even items (Groth-Marnat, 2003). How this decision is made is dependent on the characteristics of the instrument. By only administering the instrument at a single time point, the effects of time (i.e., change in the underlying construct) or multiple administrations (i.e., learning) do not affect the score. Therefore, any variation between the two scores is due to multiple constructs being measured, variation within aspects of the underlying construct, and error. There are two problems with using split-half reliability. First, longer scales have high reliability than shorter scales given the same average item correlation and, second, there are many different ways to split an instrument in half (Streiner, 2003).

[1] This description is based on the assumption that the instrument is comprised on a single scale. If the instrument is comprised of multiple scales, then the same principles apply to the scales within the instrument rather than the instrument as a whole.

The Spearman-Brown prophecy formula addresses the first problem by mathematically correcting for instrument length (Streiner, 2003). Kuder-Richardson formula 20 coefficient (KR-20) addresses the second problem by calculating the equivalent of every possible split-half reliabilities. A major limitation of KR-20 is that it is only able to calculate the reliability of dichotomous items (e.g., true/false or yes/no items).

The most general and widely used measure of internal consistency is Cronbach's coefficient alpha, α, (Streiner, 2003). Cronbach's alpha is a variation of KR-20 that is able to calculate the reliability of measures that use a multiple response format such as a one through five scale. "Alpha is defined as a portion of the scale's total variance that is attributed to a common source" (i.e., the underlying construct of interest; DeVellis, 2003, p. 31). In other words, the more the items within a scale measure single construct, the less variance is due to error (e.g., poorly worded items, participant misunderstandings or misreading, or additional constructs measured by items within the scale), and the higher the alpha coefficient. From an instrument developer's perspective in the context of research, an alpha coefficient that is too high ("much above .90"; DeVellis, 2003, p. 96) is problematic because it indicates that there's unnecessary redundancy within the scale and items can be removed from the scale (making a scale shorter and more user-friendly) without adversely impacting the scales ability to measure the underlying construct (DeVellis, 2003). However, with respect to assessment evaluation and internal consistency, a very high alpha coefficient is not problematic. Additionally, there are circumstances in which high internal consistency would not be expected. For example, a measure of exposure to traumatic events or life stressors would not be expected to have high internal consistency because there is little to suggest that somebody who has experienced the one

life stressor (death of a loved one) would have also experienced another (extended unemployment; Hayes, Nelson, & Jarrett, 1987).

Inter-rater reliability. Some instruments (e.g., semi-structured interviews, cognitive assessments, and projective instruments) require a rater or judge to interpret the response and provide a score. In these cases, *inter-rater reliability*, sometimes called *interscorer reliability* or *inter-rater agreement*, becomes a relevant metric. *Inter-rater reliability* refers to the rate of agreement between raters. When assessment instruments have low inter-rater reliability, then the variance in scores is highly influenced by error associated with the rater rather than the characteristics in the individual being measured. Ideally, the score on an instrument that relies on a rater should not be affected by the choice of one judge over another.

Inter-rater reliability can be gauged by having two raters score a set of responses by the same individual or by having two raters administer the same instrument to the same individual and evaluate the degree of agreement between the two raters (Groth-Marnat, 2003). The advantage of having two raters score a single set of responses is that this method isolates variability in scoring the individual's responses as the source of error. Having two raters administer the same instrument to the same individual introduces two other major sources of variability: (a) differences in administration procedures and (b) a learning or practice effect for the assessed individual. The former problem may be addressed via standardization of administration, discussed earlier, whereas the latter problem can be mitigated with the same strategies as discussed in the section on retest reliability.

There are a number of different ways to evaluate inter-rater reliability and the most appropriate method to use depends upon at least two factors: (a) the response options of the instrument (e.g., dichotomous, multiple, or continuous response options, duration of event) and

(b) whether item-level, time-level agreement, or overall agreement is most important (Sattler & Hoge, 2006). Among the most common ways to evaluate inter-rater reliability are the correlation between measures, intercalass correlation, and various types of item level agreement methods such as raw agreement, Cohen's kappa, and several variations of Cohen's kappa that attempt to correct some of the problems with Cohen's kappa (Sattler, 2008; Hunsley & Mash, 2008). The following paragraphs will discuss each method for assessing inter-rater reliability and how the issues listed above influence which statistical method should be used.

When the level of agreement that is of interest is the final score on an assessment instruments and the final score is either a ratio or interval scale, then inter-rater reliability can be considered in much the same way as alternate form reliability is and evaluated by correlating the two scores (alternate form reliability is discussed later in this document; Sattler & Hoge, 2006). For scores that are dichotomous, such as diagnostic assessments that determine if an individual has a given diagnosis or not, the phi coefficient is an appropriate statistical measure of inter-rater reliability. From this perspective, a correlation between the two sets of scores is calculated to indicate the degree with which the raters or administrators agree. High correlations indicate that the individual being assessed is responsible for the score on the assessment rather than the rater. Conversely, low correlations indicate that the rater or some other source of error influenced the score on the assessment to an unacceptable large degree.

Unlike using the correlation between overall scores to determine *inter-rater reliability*, each of the following methods assesses inter-rater reliability at the item level. Raw agreement refers simply to the percentage of the time that two (or more) raters agree on a score versus disagree. The raw agreement of raters using an assessment instrument is informative but, because it does not account for chance agreement, the information it provides is of limited use.

Cohen's kappa is one of the most common statistics used for calculating inter-rater reliability, is considered the standard method of assessing diagnostic agreement in psychological and medical settings, and is better than raw agreement in that it accounts for chance agreement (Uebersax, 1987; Shrout, Spitzer, & Fleiss, 1987). Cohen's kappa can be used in instances when there are two raters and both raters rate each individual and for data that is ordinal (Fleiss, 1971; Sattler & Hoge, 2006).

Concerns have been raised by many researchers about the appropriateness of using Cohen's kappa because it is overly influenced by differences in base rates even when the underlying level of agreement does not change (Uebersax, 1987). Specifically, when the base rate of an event is very high or very low, and therefore the chance agreement is very high, a given number of disagreements between two raters will represent a higher proportion of disagreements versus agreements not accounted for by chance (Shrout et al., 1987). That is, when the base rate of an event is very high or low, Cohen's kappa is limited in its ability to control for chance agreement resulting in instances where agreement between raters is high but kappa is low (e.g., Fleiss, 1971; Cicchetti & Feinstein, 1990; Lantz & Nebenzahl, 1996). To address this issue, Cicchetti and Feinstein (1990) conclude that Cohen's kappa does a sufficiently account for chance. Additionally, they recommend that Cohen's kappa always be reported along with the proportion of positive agreement (both raters agreeing that an individual possesses a given trait) and the proportion of negative agreement (both raters agreeing that an individual lacks a given trait). Their rational being that this information allows the reader to interpret Cohen's kappa in light of the raw agreement.

Additional inter-rater reliability statistics include weighted kappa and Youden's *J*. Weighted kappa statistics increasingly reduce the kappa value as the degree of disagreement

increases (Viera & Garrett, 2005). This is applicable when the ratings are made on a multipoint likert scale where not all disagreements should be considered equal. Youden's *J* is a statistic that is based entirely on the sensitivity and specificity of an instrument.

Intraclass correlations (ICC) is another statistical method for evaluating inter-rater reliability. It is appropriate when item-level agreement is important and when the data is on an interval or ratio scale. ICC is based on the ratio of within trait variance and between trait variance where the within trait variance plus the between trait variance equals the total variance for all raters across different individuals. As the ratio of between trait variance divided by total variance increases, the reliability of the raters measured by ICC increases. Because ICC is based on the correlation between variables rather than agreement alone, two raters whose scores are not the same but are correlated will receive a high ICC coefficient and therefore will be considered reliable.

Alternate form reliability. A final form of reliability to be discussed is *alternate form reliability*. *Alternate form reliability*, like its name implies, refers to the use of two, parallel forms of the same instrument at the same point in time to determine how reliably the instruments measure a trait (Groth-Marnat, 2003). Alternate form reliability is, at its core, a method for assessing content validity. If the two forms of the instrument are truly comprised of a random sampling of all the potential items that assess the construct of interest (an assumption in classical test theory) and the two forms are sufficiently large, then the two forms of the instrument will be highly correlated (DeVellis, 2003). If the correlation between the two forms is not high, then it is assumed that the items comprising the two parallel instruments include items that address a construct other than the construct of interest or the items were not randomly selected from the full pool of potential items that assess the construct.

Like *split-half reliability*, *alternate forms reliability* is not susceptible to changes in the underlying construct over time as is the case with *test-retest reliability*. Additionally, there is less of a learning effect compared to test-retest reliability because the same items are not on both assessment however some learning effects can be expected since patients can learn general strategies for responding to the items (Groth-Marnat, 2003).

Validity. While there are many different aspects of validity, validity generally refers to whether the scores on an instrument are meaningful or interpretable (Foster & Cone, 1995; Kendall & Norton-Ford, 1982). That is, a valid scale provides meaningful information about a behavior or underlying construct of interest (DeVellis, 2003). In order for a test to be considered valid, it must show a relationship with an external, independently observed event (Groth-Marnat, 2003). As opposed to reliability which is concerned with the scores on an instrument, validity is concerned with the meaning or inference made about the scores (Foster & Cone, 1995). Aspects of validity include face validity, content validity, criterion-related validity (predictive and concurrent), construct validity (convergent and discriminant validity), incremental validity, and conceptual validity with the later two being especially relevant in clinical practice (Groth-Marnat, 2003).

What is a construct? A discussion of validity should begin with a definition of the term *construct*. A *construct* is an unobservable psychological attribute or trait which is theorized to exist (Cronbach & Meehl, 1955; Smith, 2005). Psychological assessment instruments typically measure a behavior or a construct and it is important to understand the difference between the two. Behaviors and constructs are different in that behaviors can be operationally defined by an assessment instrument while a construct is distinct from the items on a measure of the construct (Foster & Cone, 1995; Cronbach & Meehl, 1955). This is relevant to our discussion of validity

in that it emphasized the need to make an inference between the scores on an instrument that measures a construct (which may be comprised of items asking about discrete behaviors) and the hypothetical construct. Validity generally refers to the accuracy of this inference and the meaning that can be construed from scores.

Since a construct cannot be directly observed and measured and since it is only theorized to exist, an important way of researching the construct and the construct validity of an instrument of the construct is by testing the strength of the nomological network in which the construct exists (Cronbach & Meehl, 1955). A nomological network is a set of relationships or predictions linking the construct of interest, associated constructs, and behaviors (Cronbach & Meehl, 1955). It is these associations with established constructs and discrete, observable behaviors that gives a construct its meaning and provides a way to evaluate the validity of the construct. Additionally, it is the relationship between the construct and real world, measureable events that gives a construct its meaning and importance (Cronbach & Meehl, 1955; APA, 1954). A construct without a nomological net is inconsequential.

Various researchers have proposed different ways of organizing validity research. For example, Foster and Cone (1995) separate validity research and aspects of validity into three parts: (1) content and face validity, (2) representational validity, and (3) elaborative validity. DeVellis (2003) divides aspects of validity into similar parts: content validity, criterion-related validity, and construct validity. The discussion of validity that follows will be organized in a similar way. Content and face validity (discussed in more detail later in this document) has the narrowest focus of the three parts. Evaluating an instrument's content and face validity can and should take place prior to administering the instrument. Content and face validity will not be included among the criteria evaluating assessment instruments for reasons discussed later in this

document. Representational validity refers to whether an instrument assesses what it claims to measure and is primarily based on convergent and divergent validity—two types of construct validity (Foster & Cone, 1995). After it is determined that an instrument assesses the construct or behavior of interest, research should focus on the instrument's elaborative validity. Elaborative validity refers to the to an instrument's ability to "understand, predict, control, or monitor changes in other phenomena" and its usefulness in practical settings (Foster & Cone, 1995, p. 250). The authors state that the principle method for evaluating elaborative validity are tests of criterion-related validity (Foster & Cone, 1995).

Content validity. *Content validity* is related to the adequacy with which the items within an instrument fully measure the various aspects of the construct which it is supposed to measure (DeVellis, 2003). Theoretically, this is accomplished by randomly sampling all of the possible items that assess the construct of interest (DeVellis, 2003). While this may be possible in some circumstances (when the construct of interest in definable in exact terms; e.g. vocabulary words assigned for homework to a particular class), in most instances that interest psychologiests, all of the possible items are not known. In reality, content validity is measured by experts in a field who review the items within the instrument and determine whether or not the items assess the relevant aspects of the construct of interest.

Face validity. *Face validity* refers to whether items, individually or as a set, appear to measure the construct they were developed to measure (DeVellis, 2003). While at first appearance, *face validity* seems to be an important consideration, DeVellis (2003) identifies three reasons why this may not be an important criteria. First, just because an item or instrument appears to measure a particular construct does not necessarily mean that it does measure that construct. Second, while it may be convenient in some circumstances for items to measure what

they appear to be measuring, this is not always a desirable trait. Consider instances in which respondents are motivated to make themselves appear in a particular way such as in forensic settings. Third, it is unclear who should judge what an item appears to be measuring. Instances are likely to come up where members of the scientific community disagree with one another on what an item appears to be measuring. Additionally, reliance on face validity would undermine the support for instruments that are based purely on criterion-related validity even if such instruments are able to provide clinicians and other decision makers with valuable information.

Construct validity. As mentioned earlier, internal consistency determines the extent to which the items in an instrument reflect a single, underlying construct irrespective of what that construct is (DeVellis, 2003). Construct validity assesses the degree to which the instrument reflects the construct of interest (DeVellis, 2003; Groth-Marnat, 2003). To establish construct validity, researchers should show both convergent validity and discriminant validity (also called divergent validity; Foster & Cone, 1995). The most straight forward step in demonstrating convergent validity is to show that the instrument correlates highly with established measures of the same construct (Foster & Cone, 1995). Frequently, however, instruments are developed to assess constructs for which an established measure does not exist. In these situations, the instrument should be compared to measures of related constructs (Foster & Cone, 1995). Based on the nomological network surrounding the construct of interest, researchers are able to make theoretically based predictions about the relationship between the score on one instrument and measures of other constructs that are included in the nomological network. For example, assume that an instrument is believed to assess construct A which is theorized to be positively correlated to construct B and negatively correlated to construct C. Convergent validity for the instrument in

question will be supported to the extent that the correlation between scores on the instrument and established instruments assessing constructs B and C are as predicted.

The description above focused on the convergence of scores on the instrument of interest and established instruments of the same or related constructs. The other major method of demonstrating construct validity is discriminant validity (Foster & Cone, 1995). Discriminant validity is demonstrated when instruments of two theoretically independent constructs are shown to be unrelated (Foster & Cone, 1995). Foster and Cone (1995) recommend that new measures should demonstrate that the absence of relationship with socioeconomic status, years of education, and social desirability.

When evaluating convergent and discriminant validity, it is important to control for variance associated with common assessment methods (Foster & Cone, 1995). As a general guideline, when assessing convergent validity, the assessment method (e.g., self-report, interview, informant-report, and observation) used by the instrument of interest should be as different as possible from the method used to collect data on related constructs (Foster & Cone, 1995). For example, if the instrument being evaluated is a self-report questionnaire, then the instruments used to measure related constructs should be a clinical interview or observational measure rather than another self-report measure. The best way of demonstrating convergent and discriminant validity is with the multitrait-multimethod matrix (MTMM matrix) with a priori predictions (Foster & Cone, 1995).

Foster and Cone (1995) note that correlation between the instrument of interest and an already established instrument should not be too high because that would indicate that the two instruments are redundant. This concern is similar to the concern about exceedingly high alpha scores in that it is a problem for the developer of the measure but is not necessarily an inherent

problem for the instrument. That is, if a new instrument correlates perfectly with a perfect measure of a construct, then they would both be good measures of the construct despite their exceptionally high correlation with one another.

Once sufficient evidence for representational validity has been demonstrated, the focus of research on an instrument can shift to elaborative validity. As mentioned earlier, elaborative validity is demonstrated primarily with criterion-related validity (Foster & Cone, 1995). A criterion can be performance on a non-test observable behavior or another instrument (typically an instrument that is more costly, longer, or otherwise more difficult to administer; Foster & Cone 1995). The essential principle in criterion-related validity is that the measure of interest is being compared to gold-standard measurement or event (DeVellis, 2003). Criterion-related validity refers to a measures ability to predict something of interest (Foster & Cone, 1995). When selecting the most appropriate criterion, the purpose of the instrument should be taken into consideration (Foster & Cone, 1995). Criterion-related validity can be divided into two forms: concurrent validity and predictive validity. Concurrent validity deals with prediction of events or performance at roughly the same time as the instrument is given whereas predictive validity deals with prediction of events or performance at a future time. For example, a measure of generalized anxiety would demonstrate concurrent validity if it accurately reflects the amount of time that an individual spends worrying in a given day at the present time (a real event) or a gold-standard instrument. Alternatively, a measure of early literacy skills that, when given to preschoolers, predicts reading ability in fifth grade would demonstrate predictive validity.

To some degree, the difference between construct validity and criterion-related validity when predicting scores on an assessment instrument is arbitrary in that both are evaluated via the relationship between two assessment instruments (DeVellis, 2003). The difference between the

two is to some degree determined by the intent of the researcher (DeVellis, 2003). When investigators are evaluating an instrument's construct validity, they want to know if the new measure acts in expected ways with respect to other constructs (i.e., instruments with demonstrated ability to measure the constructs) in the nomological net. On the other hand, when investigators are interested in criterion-related validity, they are asking a more practical question: Is the instrument being evaluated able to predict other behaviors (including performance on another assessment instrument) which are, for one reason or another, more difficult to measure?

With regard to criterion-related validity, DeVellis (2003) cautions about misinterpreting correlation with accuracy. Just because an instrument has a statistically significant correlation with a criterion event does not mean that clinicians are able to make better decisions when armed with the score on the scale. Criterion-related validity frequently deals with predicted events which are dichotomous: treat or do not treat, hospitalize or do not hospitalize. However, many instruments employ scales which are continuous. Therefore, it is tempting, and useful when done well, to divide the predictor instrument's scale into dichotomous categories such as clinically significant depression or not. Where to divide the scale is a difficult decision and there are hazards for making the cut score too high or too low (DeVellis, 2003). Assume that an instrument will be used to determine who gets treatment and high scores indicate higher need for treatment. If the threshold score is set too high, the instrument result in too many people who need treatment will not get treatment (false negatives). On the other hand, if the threshold score is too low, too many people will receive treatment who do not need it (false positives; DeVellis, 2003). Too much of either type of error reduces the instruments usefulness and, from a public health perspective, both errors are serious problems. In order to determine where the cut score

should be, investigators need to consider which of the two types of error are more costly and the purpose of the instrument.

Treatment utility. In 1987, Hayes, Nelson, & Jerrett recognized the need to organize the research on the benefits to clinical outcomes resulting from assessment and proposed the term "treatment utility of assessment" (p. 963). *Treatment utility* of psychological assessment is defined as the extent to which psychological assessments improves treatment outcome (Hayes, Nelson, & Jerrett, 1987). Other terms that have been used in the literature to refer to similar concepts include *incremental validity*, *concurrent validity*, *construct validity*, *predictive validity*, *discriminative efficiency*, and *utility* (Hayes, Nelson, & Jerrett, 1987). Based on Foster and Cone's (1995) definition of *elaborative validity* (evidence supporting an instrument's practical purpose), *treatment utility* of assessment would fit within that category. There are two basic questions asked by researcher looking at treatment utility of a given psychological assessment instrument (Nelson-Gray, 2003). First, when clinicians make treatment planning and treatment selection decisions based on a given psychological assessment, do clinical outcomes improve? Second, when clinicians are provided with treatment monitoring data based on a particular psychological assessment instrument, do clinical outcomes improve? The central thrust in both questions is the improvement in clinical outcomes. Clinicians' use of assessments with treatment utility will lead to (1) selection of more appropriate treatment and (2) appropriate and timely revision or modification of treatment (Nelson-Gray, p. 521, 2003).

With regard to assessment for the sake of diagnosis and treatment planning, *treatment utility* of assessment is the ultimate criterion (Hayes, Nelson, & Jarrett, 1987). *Reliability*, *representative validity*, and even *elaborative validity* are measured by analyzing the performance of the items within an instrument or the instrument as a whole but do not assess the degree to

which the instrument supports the ultimate goal in clinical practice—resolution of the presenting problem (Hayes, Nelson, & Jarrett, 1987).

Assessment Evaluation Systems

The discussion above represents a brief overview of the literature on the methods use to evaluate psychological assessment instruments. However, even a thorough understanding of what the characteristics of assessment instruments (standardization, reliability, validity, and utility) is insufficient to evaluate the strength of the instrument. For example, knowing the meaning and value of construct validity is insufficient knowledge for clinicians to select the most evidence-based assessment instruments available. Clinicians need a comprehensive system for evaluating assessment instruments. As discussed earlier, the Division 12 Task Force on Evidence Based Treatments (APA Division 12, 1993) has developed such a system for categorizing treatments based on the amount of empirical support for their use. An analogous system for evaluating assessment instruments could prove extremely valuable to the field by identifying criteria to consider when examining the data on the characteristics of a particular measure and condensing and organizing this information down into a manageable format. Though still underdeveloped, there have been three different initiatives that have made progress on this important front: (a) Technical Recommendations for Psychological Tests and Diagnostic Techniques later renamed Standards for Educational and Psychological Research, (b) Measures of Social Psychological Attitudes, and (c) A Guide to Assessments that Work. Each is discussed in the following section.

'Technical Recommendations for Psychological Tests and Diagnostic Techniques'. In 1954, the American Education Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) published

Technical Recommendations for Psychological Tests and Diagnostic Techniques (here afterward referred to as the Technical Manual). The purpose of the Technical Manual was to identify what information about an assessment instrument should be provided to practitioners so that they can make informed decisions about what measures to use and how to use them (AERA, APA, & NCME, 1954). The Technical Manual explicitly states that it intentionally did not include specific scores for rating assessment instruments because of the concern that different measures for different purposes have different requirements (AERA et al., 1954). It does make recommendations about the information that instrument developers should provide and suggests a standard format for presenting the information (AERA et al., 1999). Furthermore, the Technical Manual attempts to strike a balance between what information is ideally provided with what is practical given the high costs associated with conducting research (AERA et al., 1954). Therefore, the Technical Manual should be viewed as a first step and not a guide for researchers and test developers more than a half century later.

The Technical Manual is comprised of six categories of recommendations: Dissemination of Information, Interpretation, Validity, Stability, Administration and Scoring, and Scales and Norms (AERA et al., 1999). With regard to the section on validity, it identifies four types of validity: content validity, predictive validity, concurrent validity, and construct validity. As mentioned earlier, the Technical Manual does not include any specific criteria that instruments should meet.

In 1999, AERA, APA, and NCME published the Standards for Educational and Psychological Research (here afterward referred to as the Standards; AERA et al., 1999). The Standards is the fifth revision of the Technical Manual. The purpose of the Standards was to "provide criteria for evaluating tests, testing practices, and the effects of test use" (AERA et al.,

1999, p. 2). The Standards are divided into three parts: (1) test construction, evaluation, and documentation, (2) fairness of testing, and (3) testing applications. Within each part are multiple chapters, 15 in total, and within each chapter are between 11 and 27 standards. (Part one is the most relevant portion of the Standards to this study.) Similar to the Technical Manual, part one of the Standards is comprised chapters addressing: validity; reliability and errors of measurement; test development and revisions; scales norms and score comparability; test administration, scoring, and reporting; and supporting documentation for tests (AERA et al., 1999). Within the part one alone, the part of the Standards most relevant to this study, there are six domains and 123 standards.

As with the Technical Manual, the Standards do not include any guidelines for assessing the strength of an assessment instrument. For that reason and in the context of this study, it can be used to identify criterion on which an assessment instrument can or should be measured but not in developing a rating system for each criteria.

'Measures of Social Psychological Attitudes'. In 1991, Robinson, Shaver, and Wrightsman published a criteria for evaluating instruments in a book titles Measures of Social Psychological Attitudes. While their goal was to create a resource identifying and evaluating the most promising and influential measures of social attitudes across a number of different domains, they articulate 12 criteria for evaluating social psychological measures (Robinson et al., 1991). The 12 criteria they used were theoretical development/structure, pilot testing/item development, available norms, samples of respondents, inter-item correlations, coefficient alpha, factor analysis, test-retest, known groups validity, convergent validity, discriminant validity, and freedom from response set (Robinson et al., 1991).

Each criteria is on a five point scale (from zero through four: "*Exemplary*", "*Extensive*", "*Moderate*", "*Minimal*", and "*None*") with each score on each scale operationally defined (Robinson et al., 1991). However, the operational definitions vary in the strength. On the strong side, specific value ranges are listed for inter-item correlations and alpha coefficients. On the weaker side, the definition of exemplary on Available Norms is "Means and Standard Deviations for several subsamples and total sample; extensive information for each item" (Robinson et al., 1991, p. 12). This definition falls short by not specifying what information is needed to be considered "*Extensive*" and not considering the size of the sample, who is publishing the norms (author of the measure versus an independent author), whether the sample was collected through convenience sampling or is representative of the population, and how agreement or disagreement between studies impacts the strength of the norms. Despite these shortfalls, these criteria for evaluating psychological instruments represents a major step forward.

A Guide to Assessments that Work. Most recently, Hunsley and Mash (2008) edited a book titled A Guide to Assessments that Work which includes the most comprehensive psychological instrument evaluation system. Their evaluation consists of nine domains: norms, internal consistency, inter-rater reliability, test-retest reliability, content validity, construct validity, validity generalization, treatment sensitivity, and clinical utility (Hunsley & Mash, 2008). In addition to the nine domains listed above, highly recommended instruments are identified for each instrument type within each diagnostic area. The authors (2008) developed a rating scale for each domain with specific benchmarks defined. Rating options included "*Not Applicable*", "*Unavailable*", "*Adequate*", "*Good*", and "*Excellent*". Although not explicitly discussed in their rating system, instruments that fell short of the "*Adequate*" mark but for which there was evidence received a rating of "*Less than Adequate*" (Hunsley & Mash, 2008).

This rating system represents a great stride forward in that the criteria for each domain are more fully defined than previous rating systems. That being said, there is still substantial room for improvement. For example, the criteria for norms requires that the normative sample be "large" but stops short of defining what constitutes a large sample. Furthermore, here remain important aspects of a rating system that are not discussed. For example, Hunsley and Mash (2008) do not specify where the data on the measures is to be collected and over what time span the literature should be reviewed. Additionally, unlike the criteria for evaluating evidence based treatments (Chambless & Ollendick, 2001), there is no consideration for who is conducting the research on the assessment instrument. The concern is that the authors of an instrument may over state the strength of the instrument. In order to reach the highest level of support with the EBT criteria, research supporting the treatment must be published by more than one research group (Chambless & Ollendick, 2001). Additionally, the rating guidelines do not articulate how to account for contradictory evidence. For example, guidelines are not articulated for what to do if one study finds that the alpha coefficient for an instrument is in the "*Less than Adequate*" range and another is in the "*Good*" range. With regard to the evaluation of normative data, the rating system applies one rating for the measure but does not indicate for what populations there is or is not adequate normative data. One can imagine a situation where an instrument has excellent normative data on some minority populations but not the population for which a clinician wishes to use the instrument.

Statement of Purpose

Given the high demands for mental health services in the United States and the limited resources, there is justified demand that mental health services demonstrate the effectiveness of services provided to individuals. Beginning in the 1990s, great efforts were put forth to develop

criteria to evaluate the effectiveness of psychological treatment but too little emphasis was placed on evaluating assessment practices which go hand in glove with treatment. Thus far, this document has outlined the need for improved psychological practices and a number of criteria on which psychological instruments can be evaluated.

The goals of the present study were to (a) develop and pilot a tool for evaluating *Assessment Instruments* and (b) evaluate the Assessment Clinic's assessment practices using the newly developed Strength of Measure (SoM) along with other criteria described in the literature as best practices. Drawing on the knowledge gleaned from the literature and past efforts to establish criteria for evaluating *Assessment Instruments*, criteria for the SoM were established to gauge the quality of *Assessment Instruments*. These criteria consist of nine dimensions: *Readability*, *Standardization*, *Normative Data*, *Internal Consistency*, *Inter-rater Reliability*, *Test-retest Reliability*, *Construct/representative Validity*, *Criterion-related/elaborative Validity*, and *Treatment Utility*. Each is operationally defined in Appendix 1 of this document (see below). The resulting measure will then be used to evaluate the *Assessment Instruments* used as part of Assessment Reports resulting from assessments conducted and written at the CMHC's Assessment Clinic.

Hypotheses

Analysis of the data from this study tested the following hypotheses:

1. Hypothesis one states that each *Assessment Instruments* will be deemed "*Adequate*" or better on each of the nine *SoM* scales applicable to the instrument.
2. Hypothesis two states that each Report will include at least two *Assessment Instruments* deemed "*Adequate*" or better on each SoM dimension addressing the primary *Reason for Referral*.

3. Hypothesis three states that each diagnosis assigned in each Report will be supported by at least one instrument, deemed "*Adequate*" or better on each applicable SoM dimension, that specifically assesses characteristics associated with the Primary Axis I Diagnosis.
4. Hypothesis four states that each Report will include more than one *Informant* (e.g., self-report, parent report, and teacher report).
5. Hypothesis five states that each Report will include more than one Method of Assessment (e.g., interviews, paper and pencil instruments such as checklists, cognitive assessments, and projective instruments).

Method

Overview

Data for this study was collected as part of a larger program evaluation of an Assessment Clinic at a community mental health clinic (CMHC) in the mid-Atlantic region. The goals of the present study were to (a) develop and pilot a tool, the *SoM*, for evaluating *Assessment Instruments* and (b) evaluate the Assessment Clinic's practices using the newly developed tool and against other benchmarks established in the assessment literature. Data were extracted from medical records at the CMHC. Specifically, data were collected and coded from deidentified Initial Referral Forms and psychological assessment reports. The study received approval from the Virginia Commonwealth University Institutional Review Board (IRB) and the executive director of the CMHC.

Participants

Participants included every youth who received services from the CMHC's Assessment Clinic during their first year of operation—between September 1, 2008 and August 31, 2009. Since the participants in this sample include every youth for whom a report was written during a 12-month period, it is reasonable to assume that the sample used in this study closely matches the population of youth who receive services at the CMHC on an ongoing basis. The Initial Referral Forms (hereafter referred to as IRF) and the psychological assessment reports (hereafter referred to as Reports) were collected for all youth for which these documents are available. Each of the assessments and Reports for the youth included in this study were conducted by one of two clinical psychologists along with one clinical social worker. Both of the clinical psychologist were recent graduates from their doctoral programs.

During the one-year period, reports were produced for 33 youth. However, since one youth's report was based solely on a records review and the CMHC's staff did not have any direct contact with the youth or the family, this youth's report was excluded from analysis leaving a final sample 32 youth. The demographic characteristics of the youth were extracted from the text of the Reports. The mean age of the sample was 13.56 years of age (*SD* = 2.78) ranging from 8.47 to 18.39 years. Of the 32 youth, 18 were males (56.3%) and 14 were females (43.8%). The youth ranged in grade from 2nd grade to 12th grade with a mean grade of 7.03 (*SD* = 2.57). The youth's ethnicity was provided in 29 of the 32 reports with three reports not identifying the youth's ethnicity (9.4%). The sample consisted of 15 African American youth (46.9%), eight Caucasian youth (25.0%), three Multi-ethnic youth (9.4%), two Asian youth (6.3%), and one African youth (3.1%). Twenty-eight (87.5%) of the youth were referred from Department of Family Services or equivalent agencies and four (12.5%) were referred from the Judicial System. These data are also presented in Table 1 (see below).

Procedures

Identification of the sample. As noted above, staff members at the CMHC identified each youth for whom a Report was written by the Assessment Clinic in the first year of operation—between September 1, 2008 and August 31, 2009. This particular CMHC was selected because it recently developed and implemented a new assessment and diagnostic service in hopes of improving the assessment and diagnostic services provided to youth and families being served by the organization.

Table 1

Youth Demographics

Variable	N (%)	Mean (*SD*)	Minimum	Maximum
Age	32 (100%)	13.56 (2.78)	8.47	18.39
Gender	32 (100%)			
Male	18 (56.25%)			
Female	14 (43.75%			
Grade	32 (100%)	7.03 (2.57)	2	12
Youth's Ethnicity	29 (90.63%)			
African American	15 (46.88%)			
Caucasian	8 (25.0%)			
Bi/Multi Ethnic	3 (9.38)			
Asian	2 (6.25%)			
African	1 (3.13%)			
Missing	3 (9.38)			
Referral Agency	32 (100%)			
Judicial System	4 (12.50%)			
Department of Family Services	28 (87.50%)			
Assessment Period	17 (53.13%)	59.71 (18.18)	27	90
Number of Evaluation Dates	32 (100%)	5.84 (1.71)	0	9

Data Extraction and Coding

As noted above, data sources for this study were (a) Reports and (b) IRFs. Some data from this study were directly extracted from these sources and other data required various levels of coding. This section describes (a) from where the data were extracted and (b) the coding procedures. For the purposes of describing the data, variables are organized into the following categories: non-coded variables, coded variables, and calculated variables.

Non-coded variables. Non-coded variables are drawn from the IRFs and Reports for which transcription with minimal to no inference is required. The non-coded variables for this study are: (a) *Date of Birth*, (b) *Gender*, (c) *Grade in School*, (d) *Date of Report*, (e) *Number of Assessment Dates*, (f) *Ethnicity*, (g) *Referral Source*, (h) *Assessment Components*, (i) *Level of Assessment Specificity*, (j) *Primary Diagnosis*, and (k) *Date of Referral* (see Table 2 below).

Date of birth, *Gender*, *Grade in School*, and *Date of Report* are all provided in the header of the Report. Each of these variables were transcribed as written in the Report. *Number of Assessment Dates* was determined by counting the number of dates listed in the "Evaluation Dates" section of the Report header. The youth's Ethnicity and the *Referral Source* were drawn from the first paragraph of the Report. Ethnicity, when available, is clearly stated in the first section of each Report and recorded as stated in the Report. To help protect the identity of the youth, the specific agency names were not included. Instead, *Referral Source* was coded as a categorical variable with Department of Children and Families and Department of Juvenile Justice as the only two *Referral Source*s identified in this sample.

Table 2

Non-coded Variables

Variable Name	**Source of Data**	**Definition of Variable**
Date of Birth	Report	Date of Birth as reported in the Report.
Gender	Report	Gender as reported in the Report.
Grade in School	Report	Grade refers to the youth's grade in school and is defined as the grade reported in the Report.
Date of Report	Report	Date of Report is defined as the date that the Report was completed.
Number of Assessment Dates	Reports	Number of Assessment Dates refers to the number of times that a member of A&Ds assessment team met with an informant to collect information for the assessment and is determined by the number of dates listed on the Report.
Ethnicity	Report and IRF	Ethnicity refers to the youth's ethnicity as reported in the Report.
Referral Source	Report	Referral Source refers to the type of referral source that referred the youth to the Assessment Clinic (e.g., Department of Family Services or similar agency or judicial system).
Assessment Components	Report	Each assessment component that was reported in the Report was recorded. This includes meetings, structured and unstructured interviews, and specific and general assessment instruments and practices.
Diagnostic Category	Report	The first diagnosis listed for Axis I was recorded and coded into Diagnostic Categories.
Date of Referral	IRF	Date of Referral is defined as the date listed on the IRF.

In addition to these variables, the elements of the psychological assessment used in each Report were also extracted. For the purpose of this document, the term *Assessment Component*

is defined as any assessment element listed in the "Assessment Component" section of the Report and *Assessment Instrument* refers to specific and identifiable *Assessment Components* for which psychometric data is available. Forty-four *Assessment Components* were identified in the Reports. In order to record all of the elements of the assessment, each *Assessment Component* was assigned its own variable and coded *Yes* if the *Assessment Component* was present and *No* if it was not present. The Reports frequently listed "Meeting" as an *Assessment Component* and listed the individuals in attendance along with their relationship to the youth. For the purposes of this study, if a family member other than the youth was present at the meeting, "Family Interview" was coded positively. Similarly, if individuals from the youth's school or current out-of-home placement (e.g., a residential treatment center, group home, or juvenile detention center) was listed as a meeting participant, then "School Meeting" and "Placement Meeting", respectively, were coded positively.

These *Assessment Components* were organized into four categories based on how identifiable the *Assessment Component* was and the frequency with which it was used: (a) *Identifiable, High Use*; (b) *Identifiable, Low Use*; (c) *Unidentifiable, Specific Practice*; and (d) *General Practice*. The level of assessment specificity of each *Assessment Component* has implications for the level of analysis that was conducted with the component (see below). Again, the term *Assessment Instrument* refers to instruments in the *Identifiable, High Use* and *Identifiable, Low Use* categories.

Assessment Components that were frequently used (i.e., used in greater than 25% of the reports) and identifiable as a specific assessment instrument for which there is published literature and whose psychometric characteristics can be reviewed and summarized were categorized as *Identifiable, High Use*. *Assessment Components* that were similarly identifiable

but were used less than 25% of the time were in a second category, *Identifiable, Low Use*. *Assessment Components* that referred to a specific practice or procedure but were not specifically identifiable were coded as *Unidentifiable, Specific Practice*. Finally, assessment components that do not refer to a specific practice were coded as *General Practice*. A complete list of the Assessment Components, data on the frequency of use for each, and their categorization is provided in Table 3 (see below).

Table 3

Assessment Components Data

Measure	**Times used in reports**	**Level of coding**	**Informant**	**Method of Assessment**	**Reason for Referral Domains Assessed**	**Diagnosis Domains Assessed**
Rorschach	23 (71.88%)	Identifiable, High Use	Youth	Projective Measure	Mood, Psychotic	Mood, Psychotic, Personality
Trauma Symptom Checklist for Children (TSCC)	16 (50.00%)	Identifiable, High Use	Youth	Checklist/Symptom Inventory	Mood, Anxiety	Mood, Anxiety, Trauma
Teacher Report Form (TRF)	15 (46.88%)	Identifiable, High Use	Teacher	Checklist/Symptom Inventory	Mood, Inattention, Anxiety, Psychotic, Conduct,	Disruptive, Inattention, Mood, Anxiety, Psychotic,
Child Behavior Checklist (CBCL)	15 (46.88%)	Identifiable, High Use	Parent	Checklist/Symptom Inventory	Mood, Inattention, Anxiety, Psychotic, Conduct,	Disruptive, Inattention, Mood, Anxiety, Psychotic,
Weschler Intelligence Scale for Children – 4th Edition (WISC-IV)	15 (46.88%)	Identifiable, High Use	Youth	Cognitive Assessment	Impairment	Cognitive
Youth Self-Report (YSR)	15 (46.88%)	Identifiable, High Use	Youth	Checklist/Symptom Inventory	Mood, Inattention, Anxiety, Psychotic, Conduct,	Disruptive, Inattention, Mood, Anxiety, Psychotic,

Table 3 (continued)						
Roberts Apperception Test for Children: 2 (Roberts-2)	15 (46.88%)	Identifiable, High Use	Youth	Projective Measure	Mood, Anxiety, Psychotic, Conduct	Disruptive, Mood, Anxiety, Psychotic, Developmental
Children's Depression Inventory (CDI)	15 (46.88%)	Identifiable, High Use	Youth	Checklist/Symptom Inventory	Mood	Mood
Woodcock Johnson III Tests of Achievement (WJ-III: ACH)	13 (40.63%)	Identifiable, High Use	Youth	Cognitive Assessment	Impairment	Cognitive
Revised Child Manifest Anxiety Scale-2nd Edition (RCMAS-2)	12 (37.50%)	Identifiable, High Use	Youth	Checklist/Symptom Inventory	Anxiety	Anxiety
Minnesota Multiphasic Personality Inventory – Adolescent (MMPI-A)	10 (31.25%)	Identifiable, High Use	Youth	Checklist/Symptom Inventory	Mood, Anxiety, Psychotic, Conduct	Disruptive, Mood, Anxiety, Psychotic, Personality
Millon Adolescent Clinical Inventory (MACI)	7 (21.88%)	Identifiable, Low Use	Youth	Checklist/Symptom Inventory	Mood, Anxiety, Substance	Disruptive, Mood, Anxiety, Substance, Personality
UCLA-PTSD Index for DSM-IV	4 (12.50%)	Identifiable, Low Use	Youth	Checklist/Symptom Inventory	Anxiety	Anxiety, Trauma
Draw a Person	4 (12.50%)	Identifiable, Low Use	Youth	Art-based Assessment	Other	Other
Woodcock Johnson III Tests of Cognitive Abilities (WJ-III: COG)	3 (9.38%)	Identifiable, Low Use	Youth	Cognitive Assessment	Impairment	Cognitive
Test of Variable Attention (TOVA)	3 (9.38%)	Identifiable, Low Use	Youth	Behavioral Observation	Inattention	Inattention
Kinetic Family Drawing (KFD)	3 (9.38%)	Identifiable, Low Use	Youth	Art-based Assessment	Other	Other
Social Responsiveness Scale (SRS)	2 (6.25%)	Identifiable, Low Use	Parent	Checklist/Symptom Inventory	Other	Developmental
Stanford-Binet: Intellegence Scale 5th Edition (SB-5)	2 (6.25%)	Identifiable, Low Use	Youth	Cognitive Assessment	Impairment	Cognitive
Comprehensive Test of Nonverbal Intelligence (CTONI)	2 (6.25%)	Identifiable, Low Use	Youth	Cognitive Assessment	Impairment	Cognitive

Table 3 (continued)						
Screen for Child Anxiety Related Disorders (SCARED)	2 (6.25%)	Identifiable, Low Use	Youth	Checklist/Symptom Inventory	Anxiety	Anxiety
Social Communication Questionnaire (SCQ)	1 (3.13%)	Identifiable, Low Use	Parent	Checklist/Symptom Inventory	Other	Developmental
Autism Spectrum Screening Questionnaire (ASSQ)	1 (3.13%)	Identifiable, Low Use	Parent	Checklist/Symptom Inventory	Other	Developmental
Aspergers Syndrome Diagnostic Scale (ASDS)	1 (3.13%)	Identifiable, Low Use	Parent	Checklist/Symptom Inventory	Other	Developmental
Wechsler Individual Achievement Test – 2nd Edition (WIAT_II)	1 (3.13%)	Identifiable, Low Use	Youth	Cognitive Assessment	Impairment	Cognitive
Wechsler Abbreviated Scale of Intelligence (WASI)	1 (3.13%)	Identifiable, Low Use	Youth	Cognitive Assessment	Impairment	Cognitive
Vineland Adaptive Behavior Scales (VABS)	1 (3.13%)	Identifiable, Low Use	Youth	Checklist/Symptom Inventory	Impairment	Developmental
Thematic Apperception Test (TAT)	1 (3.13%)	Identifiable, Low Use	Youth	Projective Measure	Other	Personality
Child Dissociative Checklist	1 (3.13%)	Identifiable, Low Use	Youth	Checklist/Symptom Inventory	Other	Other
Sentence Completion	14 (43.75%)	Unidentifiable, Specific Practice	Youth	Projective Measure	Unidentifiable measure, not coded	Unidentifiable measure, not coded
Projective Drawing	10 (31.25%)	Unidentifiable, Specific Practice	Youth	Art-based Assessment	Unidentifiable measure, not coded	Unidentifiable measure, not coded
Attachment Story Stems	3 (9.38%)	Unidentifiable, Specific Practice	Youth	Projective Measure	Unidentifiable measure, not coded	Unidentifiable measure, not coded
Play Assessment	1 (3.13%)	Unidentifiable, Specific Practice	Clinician	Behavioral Observation	Unidentifiable measure, not coded	Unidentifiable measure, not coded
OCD Screening	1 (3.13%)	Unidentifiable, Specific Practice	Not Enough Information	Not Enough Information	Unidentifiable measure, not coded	Unidentifiable measure, not coded
Interview family	32 (100%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded

Table 3 (continued)						
Records review	31 (96.88%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Interview youth	31 (96.88%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
School meeting	29 (90.63%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Home observation	23 (71.88%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
School observation	12 (37.50%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Placement meeting	4 (12.50%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Stabilization meeting	2 (6.25%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Family Assessment and Planning Meeting (FAPT meeting)	2 (6.25%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded
Intensive Care Coordination meeting	1 (3.13%)	General Practice	Not Enough Information	Not Enough Information	General practice, not coded	General practice, not coded

Each youth's *Primary Diagnosis* was transcribed and coded by the first Axis I diagnosis listed in the Report. Rule-out diagnoses were not coded. Not including rule-out diagnoses is consistent with research conducted by Jensen and Weisz (2002) who did not include rule-out diagnoses in their primary analysis of clinic versus research based diagnosis. *Date of Referral*

was transcribed from the IRF, when available. Data on *Primary Axis I Diagnosis Category* is presented in Table 4 (see below).

Table 4

Primary Axis I Diagnosis Category

Diagnostic Category	**N (%)**	**Youth Diagnoses in each Diagnostic Category (number of instances)**
No Axis I Diagnosis	2 (6.25%)	None (2)
Youth's Impulsive/Inattentive Disorder	1 (3.13%)	Attention Deficit/Hyperactivity Disorder, NOS (1)
Youth's Mood Disorder	12 (37.50%)	Depressive Disorder, NOS (5); Major Depressive Disorder (3); Mood Disorder, NOS (3); Bipolar Disorder, NOS (1)
Youth's Anxiety Disorder	6 (18.75%)	Anxiety Disorder, NOS (5); Generalized Anxiety Disorder (1)
Youth's Trauma Related Disorder	6 (18.75%)	Posttraumatic Stress Disorder (6)
Youth's Developmental Disorder	1 (3.13%)	Asperger's Disorder (1)
Youth's Other Disorder	4 (12.50%)	Adjustment Disorders (3); Reactive Attachment Disorder (1)

Coded variables. There were 15 variables that required substantial judgment in the coding process: (a) nine *Strength of Measure* (*SoM*) variables, (b) *Informant*, (c) *Method of Assessment*, (d) *Reason for Referral*, (e) *Referral Domains Assessed*, (f) *Diagnostic Domains Assessed*, and (g) *Diagnostic Category* (see Table 5 below). Coding for each variable is discussed here.

Table 5

Coded Variables

Variable Name	**Definition of Variable**
Reason for Referral	Reason for referral refers to the reason why the youth was referred to the Assessment Clinic based on information gathered from the IRF and the Report.
Standardization	Standardization refers to the structured procedure for administering, scoring, and interpreting a psychological assessment instrument.
Normative Data	Normative Data refers to the development of descriptive data (mean and standard deviation at a minimum) about scores on an instrument based on scores from a representative sample or a known group.
Internal Consistency	Internal consistency refers to the homogeneity of items within a scale (DeVellis, 2003).
Inter-rater Reliability	Inter-rater reliability refers to the rate of agreement between raters.

Table 5 (continued)	
Test-retest Reliability	Test-retest reliability refers to the degree to which scores on a measure at one point in time are similar to scores on the same measure at a second point in time for the same individual (Kendall & Norton-Ford, 1982).
Construct/Representative Validity	Representational validity refers to whether an instrument assesses what it claims to measure and is primarily based on convergent and divergent validity—two types of construct validity (Foster & Cone, 1995).
Criterion-related/Elaborative Validity	Elaborative validity refers to the to an instrument's ability to "understand, predict, control, or monitor changes in other phenomena" and its usefulness in practical settings (Foster & Cone, 1995, p. 250).
Treatment Utility	Treatment utility of psychological assessment is defined as the extent to which psychological assessments improves treatment outcome (Hayes, Nelson, & Jerrett, 1987).
Readability	Readability refers to the reading level required to read and understand text.
Informant	Informant refers to the individual (e.g., youth, parent, teacher, other informant, or clinician) providing information that leads to the score on an assessment instrument.
Method of Assessment	Method of assessment refers to the mode, structure, or format of the assessment instrument.
Referral Domains Assessed	Referral Domains Assessed refers to the domains covered by the assessment components organized using the same options as Reason for Referral.
Diagnostic Domains Assessed	Diagnostic Domains Assessed refers to the domains covered by the assessment components organized into diagnosis categories.
Diagnosis Category	Diagnosis Category refers to the domain under which the youth's Primary Diagnosis best fits. The same domains are used for this variable as was used for Diagnostic Domains Assessed.

Strength of Measure (SoM). As discussed earlier, the nine *SoM* scales are *Readability*, *Standardization*, *Normative Data*, *Internal Consistency*, *Inter-rater Reliability*, *Test-retest Reliability*, *Construct/ Representative Validity*, *Criterion-related/Elaborative Validity*, and *Treatment Utility*. The process required to code the *SoM* variables consisted of (a) conducting a preliminary and systematic search of the literature for research addressing the instrument's psychometric strength, (b) excluding articles identified in the preliminary search that did not meet the inclusion criteria, (c) reviewing these articles and extracting relevant data that pertains to the instrument's psychometric strength, and (d) when applicable, reviewing the instrument's manual for data that pertains to the instrument's psychometric strength.

The preliminary search of the literature was conducted using the search engine PsycInfo. The search was constrained to peer-reviewed journal articles published between 1/1/1995 and

12/31/2010. Literature published prior to 1995 was excluded for two reasons. First, there was concern that the changing demographics and culture would reduce the relevance of older data. Second, there were concerns that literature that was based on the third edition of the Diagnostic and Statistical Manual would be less relevant to our current understanding of mental health disorders than literature based on the fourth edition. For each instrument, a PsycInfo search was conducted searching for the title of the instrument and common abbreviations within the title of articles and variations on the terms psychometric, reliability, validity, and factor analysis in the title or abstract of articles. The preliminary search was designed to be over-inclusive so as to not inadvertently miss important articles. After conducting the preliminary search, the title and abstract of each of the articles was reviewed by the author of this study to determine if it met additional inclusion criteria. Some articles that identified in the preliminary search were eliminated because they did not address the same version of the assessment instrument used by the Assessment Clinic (e.g., a different edition of the instrument was assessed or the same edition was translated into a different language), where not primarily psychometric in nature (e.g., an article on a treatment evaluation study that used the instrument would be excluded for the purposes of this study because it is not primarily psychometric in nature), used a sample that was not from the United States, and the sample did not consist primarily of youth. The specific search terms, number of articles identified in the preliminary search, number of articles receiving careful review, and whether or not an assessment manual was reviewed are detailed in Table 6 (see below).

To evaluate each *Assessment Instrument* on the nine *SoM* scales, each of the articles identified were reviewed to extract data relevant to the nine *SoM* scales. The *SoM* Coding

Manual (see Appendix B) was used to guide data recording. The *SoM* Coding Manual also includes directions for assessing the *Informant* and *Method of Assessment*.

Table 6

Literature Review

Assessment Instrument		
Wechsler Intelligence Scale for Children – Fourth Edition	Search Criteria	(ti="Wechsler Intelligence Scale for Children – Fourth Edition" or ti=WISC-IV) and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	16
	#of Articles Reviewed	9
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Full Scale IQ, (2) Verbal Comprehension Scale, (3) Perceptual Reasoning Scale, (4) Working Memory Scale, and (5) Processing Speed Scale
Woodcock-Johnson III: Tests of Achievement	Search Criteria	(ti="Woodcock-Johnson" or ti="Woodcock Johnson" or ti="WJ III") and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	12
	#of Articles Reviewed	6
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Reading, (2) Oral Language/Knowledge, (3) Math, (4) Written Language, and (5) Total Achievement
Minnesota Multiphasic Personality Inventory – Adolescent	Search Criteria	(ti="Minnesota Multiphasic Personality Inventory - Adolescent" or ti="MMPI-A") and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	47
	#of Articles Reviewed	20
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Hypocondriasis, (2) Depression, (3) Hysteria, (4) Psychopathic Deviate, (5) Masculinity/Femininity, (6) Paranoia, (7) Psychasthenia, (8) Schizophrenia, (9) Hypomania, and (10) Social Introversion

Table 6 (continued)		
Trauma Symptom Checklist for Children	Search Criteria	(ti="Trauma Symptom Checklist for Children" or ti="Trauma Symptom Checklist" or ti="TSCC") and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	7
	#of Articles Reviewed	3
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Anxiety, (2) Depression, (3) Anger, (4) Posttraumatic Stress, (5) Dissociation, and (6) Sexual Content
Rorschach	Search Criteria	(ti="Rorschach" and ab="exner") and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	37
	#of Articles Reviewed	18
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Color, (2) Weighted Sum of Color Responses, (3) Achromatic Color, (4) Lambda, (5) Affective Ratio, (6) Egocentricity Index, (7) Experiences Actual, (8) Percentage Good Pure Form, (9) Percentage Good Form, (10) SCZI, (11) DEPI, and (12) CDI
Roberts-2	Search Criteria	(ti="Roberts Apperception Test" or ti="Roberts-2" or ti="RAT-2") and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	1
	#of Articles Reviewed	1
	Assessment Manual Reviewed	No
	Scales Evaluated	(1) Theme Overview Scale, (2) Available Resource Scale, (3) Problem Identification Scale, (4) Resolution Scale, (5) Emotion Scale, (6) Outcome Scale, and (7) Unusual or Atypical Scale

Table 6 (continued)		
Teacher Report Form	Search Criteria	(ti="Teacher Report Form" or ti=TRF or ti="Achenbach System of Empirically Based Assessment" or ti=ASEBA) and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	14
	#of Articles Reviewed	5
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Anxious/Depressed, (2) Withdrawn/Depression, (3) Somatic Complaints, (4) Social Problems, (5) Thought Problems, (6) Attention Problems, (7) Rule-Breaking Behaviors, and (8) Aggressive Behaviors
Child Behavior Checklist	Search Criteria	(ti="Child Behavior Checklist" or ti=CBCL or ti="Achenbach System of Empirically Based Assessment" or ti=ASEBA) and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	70
	#of Articles Reviewed	32
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Anxious/Depressed, (2) Withdrawn/Depression, (3) Somatic Complaints, (4) Social Problems, (5) Thought Problems, (6) Attention Problems, (7) Rule-Breaking Behaviors, and (8) Aggressive Behaviors
Youth Self Report	Search Criteria	(ti="Youth Self Report" or ti=YSR or ti="Achenbach System of Empirically Based Assessment" or ti=ASEBA) and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	21
	#of Articles Reviewed	11
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Anxious/Depressed, (2) Withdrawn/Depression, (3) Somatic Complaints, (4) Social Problems, (5) Thought Problems, (6) Attention Problems, (7) Rule-Breaking Behaviors, and (8) Aggressive Behaviors

Table 6 (continued)		
Children's Depression Inventory	Search Criteria	(ti="Children's Depression Inventory" or ti=CDI) and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	12
	#of Articles Reviewed	0
	Assessment Manual Reviewed	Yes
	Scales Evaluated	(1) Total Score, (2) Negative Mood, (3) Interpersonal Problems, (4) Ineffectiveness, (5) Anhedonia, and (6) Negative Self-esteem
Revised Child Manifest Anxiety Scale – Second Edition	Search Criteria	(ti="Revised Child Manifest Anxiety Scale" or ti=RCMAS and (ti=psychomet* or ab=psychomet* or ti=reliability or ab=reliability or ti=validity or ab=validity or ti="factor analysis" or ab="factor analysis")
	# of Articles Identified	2
	#of Articles Reviewed	2
	Assessment Manual Reviewed	No
	Scales Evaluated	(1) Physiological Anxiety, (2) Worry, (3) Social Anxiety, and (4) Total Anxiety

Prior to evaluating an *Assessment Instrument* using the *SoM* Coding Manual the scales within the instrument on which the instrument is to be evaluated must be identified. For many *Assessment Instrument*s, the scales to include were straightforward to identify (e.g., the scales on the Wechsler Intelligence Scale for Children – Fourth Edition). However, some instruments include multiple scales and subscales and the decision regarding which scales to include is not straightforward (e.g., Minnesota Multiphasic Personality Inventory – Adolescent). For this study, the most fundamental or commonly used scales for each assessment instrument were evaluated, based whenever possible, on the instrument's manual. The evaluated scales for each assessment instrument are listed in Table 6 (see above).

Informant and method of assessment. As the names suggest, the *Informant* variable reflects the individual providing the information used for the assessment instrument and the

Method of Assessment variable reflects the method by which the information is collected. Definitions for each level of the *Informant* and *Method of Assessment* variables are provided in Tables 7 and 8 respectively (see below). Like the variables addressing the domains assessed by each *Assessment Instrument*, the *Informant* and *Method of Assessment* variables were coded by the author. If the *SoM* variables were coded for the *Assessment Instrument*, then the articles and assessment manual was used to determine the *Informant* and *Method of Assessment*. If the *SoM* variables were not coded for the assessment instrument, then these variables were coded based on information gathered from the online version of Mental Measurements and Tests Yearbook accessed through EBSCOhost. Paper and pencil based personality assessment instruments such as the Minnesota Multiphasic Personality Inventory: Adolescent and Millon Adolescent Clinical Inventory were coded as checklists/ symptom inventories and art-based assessments that are projective in nature were only coded as art-based assessments.

Table 7

Informant

Informant	**Description**
Youth	The youth provided the information leading to the scores or interpretation on the instrument.
Parent	A parent or primary caretaker (e.g., aunt, grandparent, or legal guardian) provided the information leading to the scores or interpretation on the instrument.
Sibling	A sibling of the youth provided information leading to the scores or interpretation on the instrument.
Peer	A peer of the youth provided information leading to the scores or interpretation on the instrument.
Teacher	A teacher provided the information leading to the scores or interpretation on the instrument.
Clinician	A mental health clinician provided the information leading to the scores or interpretation on the instrument.
Other	An individual familiar with the youth but not described within one of the other informant categories provided the information leading to the scores or interpretation on the instrument.

Table 8

Method of Assessment

Method of Assessment	**Description**
Checklist or symptom inventory	The method of assessment is a checklist or symptom inventory completed by the youth, parent, teacher, or other informant. (This includes self-report measures of personality such as the MMPI)
Cognitive assessment	The method of assessment is an achievement, intelligence, or other test of cognitive functioning.
Interviews	The method of assessment is an interview.
Projective measure	The method of assessment is a projective measure where the informant responds to ambiguous stimuli.
Behavioral observation	The method of assessment is behavioral observation where the clinician observes behavior and scores on the measure are the result of the observed behaviors.
Behavioral monitoring	The method of assessment is behavioral monitoring where the youth, parent, teacher, or other informant records information outside the clinician's office and as close to the time of the behavior as possible.
Art therapy assessment	The method of assessment is art-based and scores on the instrument is based on the content, form, and/or process of the art.
Objective behavioral count	The method of assessment can be considered an objective behavioral count such as number of school suspensions or arrests.
Other assessment method	The method of assessment is not captured in any other category.

Reason for referral. Each youth's *Reason for Referral* to the Assessment Clinic was coded based on information in the IRF and the first two sections of the Report—"Identifying Information and Referral Questions" and "Presenting Problems and Relevant History." See Table 9 (below) for descriptions of each *Reason for Referral*. The IRFs and Reports were independently read and the *Reason for Referral* coded by the author and one of three advanced graduate students. When a consensus was not reached between the two coders, a third coder (a faculty member in the Department of Psychology and committee member for this study) was consulted. If this coder agreed with either of the initial coders, then the agreed upon code was used. When the third coder did not agree with either of the initial coders, then the Report and IRF were jointly reviewed by the author and third coder to determine the final *Reason for Referral*. (See Appendix C for the "*Reason for Referral* Coding Sheet")

Table 9

Reason for Referral

Reason for Referral	**Description**
Youth's mood related symptoms	Referral to the Assessment Clinic due to the youth's depressive and/or manic symptoms. Characteristic diagnoses and behaviors include: Major Depressive Disorder, Bipolar, and Adjustment disorder with disturbance in mood.
Youth's attention/ impulsive problem behaviors	Referral to the Assessment Clinic due to the youth's symptoms of inattention, impulsivity, or hyperactivity. Characteristic diagnoses and behaviors include: Attention Deficit/Hyperactivity Disorder.
Youth's anxiety problems	Referral to the Assessment Clinic due to the youth's symptoms of anxiety. Characteristic diagnoses and behaviors include: Various anxiety disorder and disorders conceptually related to anxiety such as Trichotillomania, Impulse-Control Disorder, and eating disorders.
Youth's psychotic or thought disordered symptoms	Referral to the Assessment Clinic due to the youth's psychotic or thought disordered symptoms. Characteristic diagnoses and behaviors include: Schizophrenia, Delusional Disorder, and psychotic behaviors (regardless of diagnosis)
Youth's reckless or dangerous behaviors	Referral to the Assessment Clinic due to the youth's behaviors that risks their own health and safety Characteristic diagnoses and behaviors include: suicidal, self-injurious behavior, and/or risky behavior that puts the youth at risk of harm (e.g., reckless sexual behavior). Note: do not code substance use here—see below.
Youth's substance use	Referral to the Assessment Clinic due to the youth's substance use and/or behaviors associated with substance use.
Youth's conduct related behavior	Referral to the Assessment Clinic due to the youth's rule breaking, aggressive, dangerous, or destructive behaviors. Characteristic diagnoses and behaviors include: Conduct Disorder, Oppositional Defiant Disorder, fire setting, aggression, running away, and/or destructive behaviors.
Youth's impairment	Referral to the Assessment Clinic due to the youth's cognitive, language, developmental or physical impairment.
Caregiver's behaviors	Referral to the Assessment Clinic due to the caregiver's behaviors or characteristics. This includes abuse and/or abandonment (i.e., active caregiver behavior that harms the youth) or neglect, caregiver psychopathology, caregiver impairment, or caregiver inability (i.e., caregiver inaction that leads to insufficient care provided).
Caregiver incarceration	Referral to the Assessment Clinic due to the caregiver being incarcerated.
Other	Referral to the Assessment Clinic due to reasons that do not fit into any other categories.

The initial coders agreed on the *Reason for Referral* for 19 (59.4%) of the youth and disagreed for 13 (40.6%) of the youth. For the 13 youth for whom the *Reason for Referral* was not agreed upon, the third independent coder agreed with one of the initial two coders eight times

and disagreed five times. These five Reports were then jointly reviewed by the author and the faculty member to determine the *Reason for Referral*.

Referral and diagnostic domains assessed. *Referral Domains Assessed* and *Diagnosis Domains Assessed* are two variables that refer to the domains assessed by the *Assessment Instrument*s. These variables were coded by the lead author. For the 11 *Identifiable, High Use* Instruments, the variables were coded based on information gathered from the assessment manuals and journal articles. For the *Identifiable, Low Use Assessment Instruments*, the *Referral Domains Assessed* and *Diagnostic Domains Assessed* were code based on information gathered obtained from the Mental Measurements and Tests Yearbook. The difference between the *Referral Domains Assessed* and *Diagnostic Domains Assessed* variables is that the *Referral Domains Assessed* used the same domains used for the *Reason for Referral* variable while the *Diagnostic Domains Assessed* used the same domains as used for the *Diagnosis Category* variables. Tables 10 and 11 (see below) list and define each *Reason for Referral Domain Assessed* and *Diagnostic Domain Assessed* by each *Assessment Instrument*.

Table 10

Referral Domains Assessed

Referral Domains	**Description**
Youth's mood related symptoms	The assessment instrument assesses mood related symptoms (e.g., depression, mania, irritability).
Youth's attention/impulsive problem behaviors	The assessment instrument assesses symptoms of inattention, impulsivity, or hyperactivity.
Youth's anxiety problems	The assessment instrument assesses symptoms of anxiety.
Youth's psychotic or thought disordered symptoms	The assessment instrument assesses psychotic or thought disordered symptoms.
Youth's reckless or dangerous behaviors	The assessment instrument assesses behaviors that risks their own health and safety
Youth's substance use	The assessment instrument assesses substance use and/or behaviors associated with substance use.
Youth's conduct related behavior	The assessment instrument assesses rule breaking, aggressive, dangerous, or destructive behaviors.
Youth's impairment	The assessment instrument assesses cognitive, language, developmental or physical impairment.

Table 10 (continued)	
Caregiver's behaviors	The assessment instrument assesses the caregiver's behaviors or characteristics.
Caregiver incarceration	The assessment instrument assesses caregiver incarceration status.
Other	The assessment instrument assesses behaviors, symptoms, and characteristics that do not fit into any other categories.

Table 11

Diagnosis Category and Diagnosis Domains Assessed

Diagnosis Domains	**Description**
Disruptive behavior disorders	The assessment instrument assesses symptoms and behaviors associated with disruptive behavior disorders.
Inattention/impulsive Disorders	The assessment instrument assesses symptoms and behaviors associated with disorders of inattention and impulsivity.
Mood disorders	The assessment instrument assesses symptoms and behaviors associated with mood disorders.
Anxiety disorders	The assessment instrument assesses symptoms and behaviors associated with anxiety disorders .
Trauma related disorders	The assessment instrument assesses symptoms and behaviors associated with trauma disorders.
Substance use disorders	The assessment instrument assesses symptoms and behaviors associated with substance use disorders.
Psychotic disorders	The assessment instrument assesses symptoms and behaviors associated with psychotic disorders.
Personality disorders	The assessment instrument assesses symptoms and behaviors associated with personality disorders.
Developmental disorders	The assessment instrument assesses symptoms and behaviors associated with developmental disorders.
Cognitive disabilities	The assessment instrument assesses symptoms and behaviors associated with cognitive functioning.
Other disorders	The assessment instrument assesses symptoms and behaviors associated with other disorders not otherwise listed.

Diagnostic category. Each youth's *Primary Diagnosis* was recorded and then coded into *Diagnosis Categories* using the same codes used to evaluate the domains that each assessment measure evaluates (see *Diagnostic Domains Assessed* and Table 11 above). For the purposes of this study, the *Primary Diagnosis* was operationally defined as the first Axis I diagnosis listed in

the youth's Report. This coding was done by the author. Table 4 (see below) lists diagnoses coded under each *Diagnosis Category*.

Calculated variables. Two variables, *Age* and *Assessment Period*, were calculated using SPSS version 18 based on other variables. *Age*, defined as the youth's age at the time the assessment report was completed, was calculated by subtracting the *Report Date* from the youth's *Date of Birth*. This resulted in the youth's age in years. *Assessment Period*, defined as the number of days between the *Referral Date* and the *Report Date*, was calculated by subtracting the *Report Date* from the *Referral Date* (See Table 12 below). Since *Referral Date* was only available for 17 of the youth, *Assessment Period* was also only available for 17 of the youth.

Coding and data entry procedures. Data entry of non-coded variables was completed by two independent coders using PASW, version 18. In the event that the two coders disagree on the value of non-coded variables, the coders referred back to the Report and IRF and recoded the item. By requiring 100 percent agreement between the two coders, calculating and ensuring sufficiently high inter-rater reliability was irrelevant. Coded and calculated variables were entered by the first author.

Table 12

Calculated Variables

Variable Name	**Variables used to calculate new variable**	**Definition of Variable**
Age	Date of Birth Date of Report	Age refers to the youth's age at the time of the completed report.
Assessment Period	Date of Referral Date of Report	Assessment Period refers to the number of days between the youth's referral to the Assessment Clinic and the date of the Report.

Analytic Plan

The following analyses were conducted to test the hypotheses listed earlier.

1. Descriptive statistics (e.g., mean, standard deviation, range, and frequency) were provided for each type of variable listed above.
2. To evaluate the first hypothesis (i.e., that each assessment instrument would be deemed "*Adequate*" or better on each of the nine *SoM* scales applicable to the instrument), the descriptive statistics (range and frequency) for the *SoM* variable were reported. A "*Less than Adequate*" rating on any of the nine *SoM* scales by any of the *Assessment Instruments* would result in a conclusion that not all of the *Assessment Instruments* were rated "*Adequate*" or better.
3. Regarding hypothesis two (i.e., that each Report will include at least two instruments that address the *Reason for Referral*), only *Assessment Instruments* receiving a rating of "*Adequate*" or better were considered. To evaluate this hypothesis, youth were categorized by *Reason for Referral*. For each *Reason for Referral* category, the *Assessment Instruments* in each Report were reviewed to determine if at least two *Assessment Instruments* were used that address the *Reason for Referral*. The second hypothesis would have been supported if at least two *Assessment Instruments* that addressed the *Reason for Referral* that received an "*Adequate*" rating or better on each of the nine *SoM* scales were administered to each youth.
4. In evaluating hypothesis three (i.e., that each Report will include at least one instrument that addresses each diagnosis assigned), only *Assessment Instruments* that received a rating of "*Adequate*" or better were considered. To evaluate this hypothesis, youth were categorized by *Diagnostic Category*. For each *Diagnostic Category*, the *Assessment*

Instruments in each Report were reviewed to determine if at least one *Assessment Instrument* was used that addresses the *Diagnostic Category*. The third hypothesis would be supported if an *Assessment Instrument* addressing the youth's *Diagnostic Category* and receiving an "*Adequate*" rating or better on each of the nine *SoM* scales was administered to each youth.

5. To evaluate hypothesis four (i.e., that each Report will include more than one informant, for instance, self-report, parent report, teacher report, and clinician report), the *Informant* variable was evaluated for each youth. Mean, standard deviation, and frequency of the *Informant*s variable was reported. This hypothesis would be supported only if each Report included data from more than one *Informant*.
6. To evaluate hypothesis five (i.e., that each Report will include more than one method of assessment, for instance, interviews, paper and pencil instruments such as checklists, cognitive assessments, and projective instruments), the *Method of Assessment* variable was evaluated for each youth. Mean, standard deviation, and frequency of *Methods of Assessment* was reported. This hypothesis would be supported if each Report included data collected via two or more *Methods of Assessment*.

Results

Overview

Results from the study are presented as follows. First, preliminary data analyses are presented, including descriptive statistics for the major variables in this study. Second, the primary analyses of the five hypotheses in this study are presented. Finally, post-hoc analyses, further investigating the data related to hypotheses one, two, and three, are presented.

Preliminary Analyses

This section provides descriptive statistics (i.e., mean, standard deviation, minimum, maximum, frequency as applicable) for each of the major variables for this study. These variables are: (a) *Reason for Referral*, (b) Report Level Variables, (c) Assessment Level Variables (*SoM*, *Informant*, *Method of Assessment*, *Referral Domains Assessed*, *Diagnostic Domains Assessed*), and (d) *Diagnostic Category*. All of the following statistics were calculated using PASW, version 18. Demographic data on the youth are described earlier (see Table 1 above).

Assessment level variables. This section discusses the following variables: *Assessment Components* (discussed in greater detail than above), *SoM*, *Informant*, *Method of Assessment*, *Referral Domains Assessed*, and *Diagnosis Domains Assessed*. The *Assessment Components* mentioned earlier can be divided by how identifiable the instrument or practice is and its frequency of use: *Identifiable/High Use*, *Identifiable/Low Use*, *Unidentifiable/Specific Practice*, and *General Practice*. Each *Assessment Component* and its level of identifiablity and frequency of use are presented in Table 3 (see above). Descriptive statistic for the Assessment Level variables are presented in Table 13 (see below).

Strength of Measure (SoM). As noted above, the *Strength of Measure* (*SoM*) variable was calculated for all of the *Identifiable/High Use Assessment Components*. *Readability* and *Standardization* were grouped together as aspects on an instrument addressed prior to administering the instrument to a single individual. The *Normative Data* category was not categorized with any other dimensions. *Internal Consistency*, *Inter-rater Reliability*, and *Test-retest Reliability* were grouped together since each addresses a component of reliability. *Construct/Representative Validity* and *Criterion-Related/Elaborative Validity* were grouped

together since each addresses a component of validity. Treatment Utility was also not grouped with any other dimensions since it is the only dimension addressing utility.

Table 13

Assessment Level Variables

Variable (Number of reports on which data is based)	Mean (*SD*)	Minimum	Maximum
Assessment Period (17)	59.71 (18.18)	27	90
Number of Evaluation Dates (32)	5.84 (1.71)	2	9
Number of *Assessment Components* - Total (32)	12.47 (3.50)	2	18
Number of *Assessment Components* - Identifiable/High Frequency (32)	5.12 (2.00)	0	8
Number of *Assessment Components* - Identifiable/Low Frequency (32)	1.25 (1.08)	0	4
Number of *Assessment Components* - Unidentifiable/Specific Practice (32)	.91 (.68)	0	2
Number of *Assessment Components* - General Practice (32)	5.19 (1.28)	2	9
Number of *Assessment Components* - All Identifiable (32)	6.38 (2.49)	0	11

After reviewing the literature and coding the nine *SoM* domains for each instrument, *Readability* was deemed applicable for 5 of the instruments and "*Not Applicable*" for six of the instruments. *Readability* was considered "*Not Applicable*" if the youth was not required to read in order to complete the instrument. In the case of self-report instruments where data were presented supporting the use of an audio version or the clinician reading the text to the youth, *Readability* was coded as "*Not Applicable*". However, if the assessment manual or other documents states that reading the measure out loud to the youth is acceptable but did not provide specific data supporting this claim, as is the case for the Children's Depression Inventory (CDI), then *Readability* was deemed Applicable and was coded based on the *Readability* data provided for the instrument. Overall, for the five instruments for which *Readability* was applicable, the mean rating was 2.2 (*SD*=1.30), ranging from "*Less than Adequate*" (0) to "*Excellent*" (3).

Standardization was a unique dimension for a couple of reasons. First it is the only dimension which was not coded on a four point scale ranging from "*Less than Adequate*" to "*Excellent*" plus "*Not Applicable*". The scale for *Standardization* included only two points, "*Less than Adequate*" (scored as zero) and "*Excellent*" (scored as three) plus "*Not Applicable*". Second, it is the only dimension for which none of the 11 instruments was coded as "*Less than Adequate*". Since *Standardization* is only a two point scale and none of the instruments were coded as "*Less than Adequate*", the mean rating was 3.0 with no variance.

Normative Data were determined to be a relevant dimension for each of the 11 instruments evaluated. The mean value for *Normative Data* were 1.64 (*SD*=1.21) ranging from "*Less than Adequate*" to "*Excellent*". Eight of the instruments were rated as having "*Adequate*" *Normative Data* or better.

The mean value for *Internal Consistency* was 0.40 (*SD*=0.52) ranging from "*Less than Adequate*" to "*Adequate*" with one instrument, the Roberts-2, coded as "*Not Applicable*". *Internal Consistency* was determined to be "*Not Applicable*" for the Roberts-2 because the scores on the individual constructs within the Roberts-2 do not necessarily measure different aspects of the same general construct and the scores on the constructs are not combined to create an overall score for the instrument. Four of the instruments were rated as having "*Adequate*" *Internal Consistency* and none of the instruments were coded above "*Adequate*". The mean value for *Inter-rater Reliability* was 0.25 (*SD*=0.50) ranging from "*Less than Adequate*" to "*Adequate*" with seven instruments being coded as "*Not Applicable*". As per the coding criteria set for the *Inter-rater Reliability* dimension (See Table 8), *Inter-rater Reliability* is only applicable when the instrument is completed by a trained observer or interviewer. Self-report, teacher-report, or caregiver-report measures are thus not subject to inter-rater reliability. Only one instrument, the

Wechsler Intelligence Scale for Children – Fourth Edition, was rated as having "*Adequate*" *Inter-rater Reliability*. The mean value for Test-Retest Reliability was 0.18 (*SD*=0.40) ranging from "*Less than Adequate*" to "*Adequate*". All 11 instruments were coded on the Test-Retest Reliability dimension but only two instruments, the Wechsler Intelligence Scale for Children – Fourth Edition and the Child Behavior Checklist, were coded as "*Adequate*".

Construct/Representative Validity and *Criterion-Related/Elaborative Validity* are two forms of validity. The mean value for *Construct/Representative Validity* was 0.09 (*SD*=0.30) ranging from "*Less than Adequate*" to "*Adequate*". Only one instrument, the Wechsler Intelligence Scale for Children – Fourth Edition, was coded as "*Adequate*" on the *Construct/Representative Validity* dimension. Each instrument was coded as "*Less than Adequate*" on the *Criterion-Related/Elaborative Validity* dimension of the *SoM*. None of the instruments were coded as "*Not Applicable*" on either of the validity dimensions of the *SoM*.

The final dimension on the *SoM* is *Treatment Utility*. *Treatment Utility* was coded for each of the 11 instruments evaluated with the *SoM*, however, all of the instruments were coded as "*Less than Adequate*". All of the *SoM* codes for each instrument are presented in Table 14 (see below). Also in Table 14, the mean and standard deviation for each *SoM* dimension across instruments, each instrument across dimensions, and each type of dimension across instruments are displayed.

Informant and method of assessment. *Informant* and *Method of Assessment* were coded for each *Assessment Component* in the *Identifiable/High Use* and *Identifiable/Low Use* categories. Thus, 29 *Assessment Components* were coded.

The youth was the *Informant* for 23 of the instruments, a parent was the informant for five of the instruments, and a teacher was the informant for one of the instruments. No instruments were based on a sibling, peer, or other informant.

Five different methods of assessment were used across the 32 Reports: (a) checklists/symptom inventories, (b) cognitive assessments, (c) projective assessments, (d) art-based assessments, and (e) behavioral observations. Sixteen of the 29 instruments were coded as Checklist/Symptom Inventories, seven were Cognitive Assessments, three were Projective Measures, two were Art-based Assessments, and one was a Behavioral Observational Measure. These data are displayed in Table 3.

Referral and diagnostic domains assessed. Each Assessment Component was coded to determine which domains were assessed from the list of Referral Domains Assessed and Diagnostic Domains Assessed. Like the Informant and Method of Assessment variables, only the Identifiable/High Use and Identifiable/Low Use were coded for the Referral and Diagnostic Domains Assessed. These codes were not mutually exclusive such that a single assessment instrument could assess multiple domains. Additionally, if an assessment measure did not define the domain or domains that were assessed, it was as assessing any domains. These data are presented in Table 3 (see above) and 15 (see below). Table 3 lists each assessment instrument and the domains assessed by the instrument. Table 15 lists the Reason for Referral and Diagnosis Domains and each instrument that assesses that domain.

Table 14

SoM Results

Instrument (times administered)	Readability	Standardization	Normative Data	Internal Consistency	Inter-rater Reliability	Test-retest Reliability	Construct/ Representative Validity	Criterion Related/Elaborative Validity	Treatment Utility	Mean for the Instrument (*SD*)
Rorschach (23)	NA	3	0	0	0	0	0	0	0	0.38 (1.06)
TSCC (16)	NA	3	2	1	NA	0	0	0	0	0.86 (1.21)
WISC-IV (15)	NA	3	3	0	1	1	1	0	0	1.13 (1.25)
Roberts-2 (15)	NA	3	0	NA	0	0	0	0	0	0.43 (1.13)
TRF (15)	3	3	2	0	NA	0	0	0	0	1 (1.41)
CBCL (15)	3	3	2	1	NA	1	0	0	0	1.25 (1.28)
YSR (15)	3	3	2	1	NA	0	0	0	0	1.13 (1.36)
CDI (15)	0	3	3	0	NA	0	0	0	0	0.75 (1.39)
WJ III Achievement (13)	NA	3	3	0	0	0	0	0	0	0.75 (1.39)
RCMAS-2 (12)	NA	3	1	1	NA	0	0	0	0	0.71 (1.11)
MMPI-A (10)	2	3	0	0	NA	0	0	0	0	0.63 (1.19)
Mean Rating (*SD*)	2.20 (1.30)	3.00 (NV)	1.64 (1.21)	0.40 (0.52)	0.25 (0.50)	0.18 (0.40)	0.09 (0.30)	0 (NV)	0 (NV)	0.82 (0.29)
Mean across stage of measure development/ evaluation (*SD*)	2.75 (0.77)		1.64 (1.21)	0.28 (0.46)			0.046 (0.21)		0 (NV)	

NV=No Variance; SD=Standard Deviation

Table 15

Reason for Referral and Diagnostic Domains Assessed

Domain	**Number of Assessments**	**Measures assessing each domain**
Referral Domain Assessed-Mood	9	YSR, CBCL, TRF, TSCC, CDI, MMPI-A, Rorschach, MCMI, Roberts-2
Referral Domain Assessed-Attention	4	YSR, CBCL, TRF,TOVA
Referral Domain Assessed-Anxiety	10	YSR, CBCL, TRF, TSCC, MMPI-A, MCMI, RCMAS-2, UCLA-PTSD, SCARED, Roberts-2
Referral Domain Assessed-Psychotic	7	YSR, CBCL, TRF, MMPI-A, MCMI, Rorschach, Roberts-2
Referral Domain Assessed-Reckless	0	NA
Referral Domain Assessed-Substance	1	MACI
Referral Domain Assessed-Conduct	5	YSR, CBCL, TRF, MMPI-A, Roberts-2
Referral Domain Assessed-Impairment	8	WISC-IV, WJ III:ACH, VABS, WJ III:COG, SB 5th, CTONI, WIAT-II, WASI
Referral Domain Assessed-Caregiver Behavior	0	NA
Referral Domain Assessed-Caregiver Incarcerated	0	NA
Referral Domain Assessed-Other	8	Draw-A-Person, Kinetic Family Drawing, SRS, SCQ, ASSQ, ASDS, TAT, CDC
Diagnostic Domain Assessed-Disruptive Dx	6	YSR, CBCL, TRF, MMPI-A, MCMI, Roberts-2
Diagnostic Domain Assessed-Inattention-Dx	4	YSR, CBCL, TRF,TOVA
Diagnostic Domain Assessed-Mood-Dx	9	YSR, CBCL, TRF, TSCC, CDI, MMPI-A, Rorschach, MCMI, Roberts-2
Diagnostic Domain Assessed-Anxiety-Dx	10	YSR, CBCL, TRF, TSCC, MMPI-A, MCMI, RCMAS-2, UCLA-PTSD, SCARED, Roberts-2
Diagnostic Domain Assessed-Trauma-Dx	2	TSCC, UCLA-PTSD
Diagnostic Domain Assessed-Substance-Dx	1	MACI
Diagnostic Domain Assessed-Psychotic-Dx	7	YSR, CBCL, TRF, MMPI-A, MCMI, Rorschach, Roberts-2
Diagnostic Domain Assessed-Personality-Dx	4	MMPI-A, Rorschach, MACI, TAT
Diagnostic Domain Assessed-Developmental-Dx	6	SRS, SCQ, ASSQ, ASDS, VABS, Roberts-s
Diagnostic Domain Assessed-Cognitive-Dx	7	WISC-IV, WJ III:ACH, WJ III:COG, SB 5th, CTONI, WIAT-II, WASI
Diagnostic Domain Assessed-Other	1	Draw-A-Person, Kinetic Family Drawing, CDC

Diagnosis. A total of 14 different Primary, Axis I Diagnoses were assigned to the 32 youth in this study. These data are presented in Table 4 (see above) along with the Diagnostic Code under which they were categorized.

Primary Analyses

Hypothesis 1: Assessment instrument adequacy. Hypothesis one stated that each assessment instrument would be deemed "*Adequate*" or better on each of the nine dimensions applicable to the instrument. This hypothesis was rejected since no instrument was coded as "*Adequate*" or better on each of the nine *SoM* dimensions. Moreover, no instrument was found to be "*Adequate*" on either the *Criterion-Related/Elaborative Validity* dimension or the Treatment Utility Dimension. On the positive side, every instrument evaluated was coded as "*Excellent*" on the *Standardization* dimension.

Hypothesis 2: Multiple instruments assessing reason for referral. Hypothesis two stated that each Report will include at least two instruments deemed "Adequate" or better that assess the dimension identified as the primary reason for referral. Because hypothesis one was rejected, hypothesis two was also rejected.

Hypothesis 3: Instrument assessing assigned diagnosis. Hypothesis three stated that each *Primary Diagnosis* assigned in each *Report* would be supported by at least one instrument, deemed "*Adequate*" or better on each applicable dimension, which specifically assesses the diagnostic category associated with the youth's *Primary Diagnosis*. Like hypothesis two, hypothesis three was rejected because there are no instruments deemed "Adequate" or better on each of the *SoM* dimensions.

Hypothesis 4: Multiple informants. Hypothesis four stated that each assessment and Report will include more than one informant (e.g., self-report, parent report, and teacher report). This hypothesis was evaluated based only on *Identifiable/High Use* and *Identifiable/Low Use Assessment Components*. This hypothesis was rejected. Of the 32 reports, the mean number of informants was 1.91 ($SD = 0.78$), ranging from zero to three. One of the Reports was based on

zero *Informants*, eight were based on only one *Informant*, 16 were based on two *Informants*, and seven were based on three *Informants*. The report that included zero *Informants* was based on two assessment components: Family Interview and Records Review. In all, 23 of the 32 Reports (71.88%) included two or more *Informants* thereby meeting the criterion set forth by hypothesis four.

Hypothesis 5: Multiple methods. Hypothesis five stated that each assessment and Report will include more than one method of assessment (e.g., interviews, paper and pencil instruments such as checklists, cognitive assessments, and projective instruments). This hypothesis was evaluated based only on *Identifiable/High Use* and *Identifiable/Low Use* assessment components. This hypothesis was rejected. Of the 32 reports, the mean number of assessment methods used was 2.72 (SD = 0.89), ranging from zero to four. One of the reports were based on zero assessment methods, one was based on only one assessment methods, nine were based on two assessment methods, 16 were based on three assessment methods, and five were based on four assessment methods. Though this hypothesis was not supported, a total of 30 of the 32 Reports (93.7%) included two or more *Methods of Assessment*.

Post-hoc Tests

The following section includes further analysis of data. These analyses are: (1) evaluating the relationship between the sum of the *SoM* scores and the frequency of use for the 11 Identifiable/High use assessment instruments, (2) reanalysis of the data associated with hypothesis two with less stringent criteria, (3) exploring whether youth from each *Reason for Referral* Domain were administered more assessment instruments assessing the *Reason for Referral* Domain than youth referred for other reasons, (4) reanalysis of the data associated with hypothesis three with less stringent criteria, and (5) evaluating whether youth from each *Primary*

Diagnosis Domain were administered more assessment instruments assessing the Diagnosis Domain than youth with other primary diagnoses.

Post-hoc analysis of assessment instrument adequacy. Hypothesis one evaluated the *SoM* of the 11 most frequently used instruments among the 32 Reports and found that none of the instruments were rated "*Adequate*" or better on each of the *SoM* dimensions. In post-hoc analysis, the relationship between an instrument's mean *SoM* rating across the nine *SoM* dimensions and its frequency of use in the 32 Reports was evaluated. A Pearson Correlation between the mean *SoM* rating and frequency of the 11 *Identifiable/High Use Assessment Component* was conducted. The results indicate that there is a non-significant relationship between the frequency of use and mean *SoM* rating: $r = -.23$, $p = 0.50$. Though the finding was not significant, the magnitude of the correlation coefficient is large enough that, had more assessment instruments (i.e., approximately 80 instruments) been coded on the *SoM* dimensions and the correlation coefficient maintained its magnitude, then the relationship could be significant. It is noteworthy and concerning that the sign of the correlation between *SoM* scores and frequency of use was negative indicating that, among the most frequently used instruments, the instruments with greater psychometric strength were used with less frequency than the instruments with less psychometric strength.

Post-hoc analysis of multiple instruments assessing reason for referral. Despite each youth not being assessed by at least two *Assessment Instruments* addressing the youth's *Reason for Referral*, meaningful data were still available regarding the number of assessment instruments used to address the *Reason for Referral* Domain for which a youth was referred to the Assessment Clinic. Hypothesis two was found to be false simply because none of the *Assessment Components* met the "*Adequate*" or better level on the nine *SoM* dimensions (see

Table 14 above). However, it is still possible that at least two assessment instruments were used to evaluate the primary reason for referral. To test this less rigorous criteria, the same analyses used to assess hypothesis two was conducted, removing the stipulation that only *Assessment Instruments* receiving an "*Adequate*" or better rating on each of the *SoM* dimensions be used. Additionally, the data can be analyzed to determine if youth from a particular *Reason for Referral* received more assessment measures addressing that domain than youth who were referred for other reasons.

Table 15 (see above) displays data on the number of assessment instruments addressing each *Reason for Referral* Domain Assessed. Table 16 (see below) displays data on the frequency of each *Reason for Referral* and the number of *Identifiable Assessment Instruments* assessing the youth *Reason for Referral* that were administered.

These data show that, of the 30 youth who were coded as having a specific *Reason for Referral* (i.e., not coded as Other *Reason for Referral*), 17 (56.67%) received at least two *Assessment Instruments* addressing the youth's *Reason for Referral* while six (20.00%) received one and seven (23.33%) received zero assessments addressing their *Reason for Referral*. Therefore, even this less stringent standard was not met.

These data show that, of the 30 youth who were coded as having a specific *Reason for Referral* (i.e., not coded as Other *Reason for Referral*), 17 (56.67%) received at least two *Assessment Instruments* addressing the youth's *Reason for Referral* while six (20.00%) received one and seven (23.33%) received zero assessments addressing their *Reason for Referral*. Therefore, even this less stringent standard was not met.

Table 16

Reason for Referral Analyses

	N	# of measures (all specific measures; N = 29) that address Reason for Referral	Included 3 or more	Included 2	Included 1	Included 0	Mean for only specific Referral Reports (*SD*)	Mean for Reports of other Reason for Referrals (*SD*)	T-Test
Youth's Conduct	19	5	10	5	4	0	2.37 (0.89)	1.92 (1.19)	NS: t(30) = 0.51, p = 0.61, r = 0.22
Youth's Impairment	5	8	1	1	2	1	1.40 (1.14)	1.15 (1.00)	NS: t(30) = 1.21, p = 0.25, r = .09
Caregiver's Behavior	5	0	0	0	0	5	0 (NV)	0 (NV)	--
Youth's Reckless Behavior	1	0	0	0	0	1	0 (NV)	0 (NV)	--
Other	2	--	--	--	--	--		--	--
Total (not including Other)	30	13	11 (36.67 %)	6 (20.00 %)	6 (20.00 %)	7 (23.33 %)			

To evaluate whether youth in each *Reason for Referral* category received more assessment instruments assessing that *Reason for Referral* Domain than youth in other *Reason for Referral* categories, a series of independent samples *t*-tests were conducted with each level of *Reason for Referral* Domain as the independent variables and the number of assessment instruments administered from the corresponding *Reason for Referral Domains Assessed* as the dependent variable. To minimize the risk of Type I errors associated with conducting multiple analyses with the same set of data, a modified Bonferroni correction was used (see Holm, 1979). In this procedure, the alpha level for the set of *t*-tests is set at .05 divided by the number of analyses conducted, in this case, .05/2 or .025, for the two analyses associated with *Reason for Referral*. If, however, the *t*-tests with the lowest *p*-value in the series is significant at the more

conservative alpha level, then the alpha level for the *t*-test with the second lowest *p*-value is set at .05 divided by the number of analyses remaining, in this case one. On the other hand, once one of the *t*-tests is not significant, the remaining analyses are deemed not significant.

First, a *t*-test was conducted with referred for *Youth's Impairment* as the independent variable and number of assessment instruments coded as *Youth's Impairment Domain Assessed* as the dependent variable. On average, youth who were referred to the Assessment Clinic for impairment issues received 1.40 (*SD* = 1.14) *Assessment Instruments* coded as *Youth's Impairment Domain Assessed* while youth without this *Reason for Referral* received 1.15 (*SD* = 1.00) assessments in this domain, a difference that was not statistically significant at the *p*=.025 level, $t(30) = 1.21$, $p = 0.25$.

As a result of the first t-test being not significant, the remaining analysis in this series must also be considered not significant. The second, *t*-test was conducted with *Referred for Youth's Conduct Behavior* as the independent variable and number of assessment instruments coded as *Youth's Conduct Behavior Domain Assessed* as the dependent variable. On average, youth who were referred to the Assessment Clinic for conduct behavior issues received 2.37 (*SD* = 0.89) *Assessment Instruments* coded as *Youth's Conduct Behavior Domain Assessed* while youth without this *Reason for Referral* received 1.92 (*SD* = 1.19) *Assessment Instruments* in this domain, $t(30) = 0.51$, $p = 0.61$.

T-tests addressing *Caregiver Behavior* and *Youth's Reckless Behavior* were not conducted since there were no *Assessment Instruments* included that specifically addressed these issues. Additionally, an analysis was not conducted for Other *Reason for Referral* because this is not a unified construct. These data indicate that there was not a significant difference between

the number of *Assessment Instruments* assessing a particular *Reason for Referral* Domain given to youth referred for that domain compared with youth referred for other reasons.

Post-hoc analysis of instruments assessing primary diagnosis domain. This section further evaluates data related to hypothesis three: that each primary diagnosis assigned in each Report would be supported by at least one *Assessment Instrument*, deemed "*Adequate*" or better on each applicable dimension, that specifically assesses the diagnostic category associated with the youth's primary axis one diagnosis. As in the case of the hypothesis addressing *Reason for Referral*, the hypothesis addressing *Youth's Primary Diagnosis* was also not supported simply because there were no *Assessment Instruments* deemed "*Adequate*" or better. However, despite this, meaningful data is still available regarding the *Assessment Instruments* administered and the diagnoses assigned.

First, the data were reviewed to determine if at least one *Assessment Instrument* was administered that assessed each primary diagnosis. Table 17 displays the frequency with which youth receiving a primary diagnosis in each *Diagnostic Domain* and the number of *Assessment Instruments* administered that address *Diagnostic Domain*. They show that each of the 26 youth whose primary Axis I diagnosis fit into a defined *Diagnostic Domain*, all 26 received at least one assessment instrument that was coded as assessing that *Diagnostic Domain*.

Next, to evaluate whether youth who received a *Primary Diagnosis* in a particular *Diagnostic Domain* were administered more *Assessment Instruments* addressing that domain than youth who were not assigned a *Primary Diagnosis* in that domain, a series of independent samples *t*-tests were conducted with each level of *Diagnostic Domain* as the independent variables and the number of *Assessment Instruments* administered from the corresponding *Diagnostic Domain Assessed* as the dependent variable. A series of t-tests were used to evaluate

the *Youth's Anxiety Disorder Domain* and the *Youth's Mood Disorder Domain*. Analyses of *Youth's Trauma Diagnosis Domain*, *Youth's Inattention/Impulsivity Domain*, and *Youth's Developmental Impairment Domain* were not conducted since there was no variance for the number of *Assessment Instruments* administered for these three groups of youth. For these analyses, the same modified Bonferroni correction for Type I errors associated with multiple analyses was made for the analyses of the *Primary Diagnosis Domain*.

First, a *t*-test was conducted with primary *Anxiety Disorder Diagnosis* as the independent variable and number of *Assessment Instruments* coded as *Anxiety Disorder Domain Assessed* as the dependent variable. On average, youth whose primary Axis I diagnosis was in the *Anxiety Disorder Domain* received 3.83 (*SD* = 1.47) *Assessment Instruments* coded as *Anxiety Disorder Domain Assessed* while youth without these diagnoses received 3.38 (*SD* = 1.44) *Assessment Instruments* in this domain, a difference that was not statistically significant at the modified p-value of .025, $t(30) = .68$, $p = 0.50$.

As a result of the first t-test being not significant, the remaining analysis in this series must also be considered not significant. The second *t*-test was conducted with *Primary Mood Disorder Diagnosis* as the independent variable and number of *Assessment Instruments* coded as *Mood Disorder Domain Assessed* as the dependent variable. On average, youth whose primary Axis I diagnosis was in the *Mood Disorder Domain* received 4.33 (*SD* = 1.72) Assessment Instruments coded as *Mood Disorder Domain Assessed* while youth without these diagnoses received 3.95 (*SD* = 1.67) *Assessment Instruments* in this domain, $t(30) = .62$, $p = 0.54$.

In sum, these results indicate that youth with *Primary Axis I Mood* and *Anxiety Disorders* did not receive a statistically different number of *Assessment Instruments* addressing each respective *Diagnosis Domain* than youth without these diagnoses. These data along with the

data for *Youth's Trauma Diagnosis Domain*, *Youth's Inattention/Impulsivity Domain*, and *Youth's Developmental Impairment Domain* are presented in Table 17 (see below).

Table 17

Diagnosis Category Analysis

	N	# of measures (all specific measures; N = 29) that address Diagnosis	Included 3 or more	Included 2	Included 1	Included 0	Mean for only specific Diagnosis Reports	Mean of all reports	T-Test
Youth's Mood Disorder	12	8	10	1	1	0	4.33 (1.72)	3.95 (1.67)	NS: t(30) = .62, p = .54, r = 0.11
Youth's Anxiety Disorder	6	8	5	1	0	0	3.83 (1.47)	3.38 (1.44)	NS: t(30) = .68, p = .50, r = 0.12
Youth's Trauma Diagnosis	6	2	0	0	6	0	1 (NV)	0.54 (0.51)	--
Other	4	0	0	0	0	4	--	--	--
None	2	--	--	--	--	--	--	--	--
Youth's Inattention/ Impulsivity	1	4	1	0	0	0	3 (NV)	1.45 (0.99)	--
Youth's Developmental Impairment	1	5	1	0	0	0	3 (NV)	0.58 (0.72)	--
Total (not including Other and None)	26	27	17 (65.38 %)	2 (7.69 %)	7 (26.92 %)	0 (0.00 %)			

NV = No Variance

Discussion

The aims of this study were to (a) develop and pilot a standardized tool and method for evaluating the degree to which assessment instruments are supported by the literature (i.e., the *SoM*) and (b) evaluate the Assessment Clinic's practices using the *SoM* and against other criteria outlined in the literature. Related to these aims, five specific hypotheses were developed and tested. While none of the hypotheses were supported, this study provides important lessons

related to the evaluation of general practices used to conduct psychological assessments in a clinical service context.

The sections to follow discuss the major findings from this study. In brief, this study contributed to the field by (a) developing a tool to evaluate *Assessment Instruments*, (b) finding that many of the *Assessment Instruments* widely used and respected by researcher and clinicians have weaker psychometric strength than previously thought, and (c) raising concerns about the general assessment practices used in community assessment contexts. In addition to discussing these findings in greater detail, the directions for future research related to the *SoM* in particular, the evaluation of *Assessment Instruments* and assessment practices in general are discussed. Additionally, the limitations of this study along with ways to address these weaknesses are also discussed.

Strength of Measure and Evaluation of Assessment Instruments

One of the primary goals of this study was the development and demonstrated use of a tool to evaluate the psychometric strength of assessment instruments—the *Strength of Measure (SoM)* scale. The *SoM* is a multi-dimensional, criterion-based assessment of the psychometric strength of assessment instruments with ratings, ranging from "*Less than Adequate*" to "*Excellent*," corresponding to generally accepted, but poorly defined, criteria established by the field. The dimensions of the *SoM* and the specific criteria were developed based on more than a half century of research addressing assessment theory, psychometrics, and assessment practices (e.g., Cronbach & Meehl, 1955; AERA et al., 1954; Robinson et al., 1991; Foster & Cone, 1995; Mash & Hunsley, 2005; Hunsley & Mash, 2008).

Past efforts to develop assessment evaluations systems (e.g., Hunsley & Mash, 2008) have relied on expert ratings and included phrases such as "a preponderance of evidence" that

lack operational definitions. This system, due to its dependence on expert ratings, is problematic because it does not lend itself to independent replication (a cornerstone of science) of the *Assessment Instrument* evaluations and it fails to communicate to the research community the exact criteria used to evaluate the *Assessment Instruments*. The *SoM* was designed to avoid dependence on expert raters and to use clearly defined criteria so that evaluation of *Assessment Instruments* would be transparent and replicable. The *SoM* was modeled in part on the criteria for evidence-based treatments (EBTs) published by Division 12 of the American Psychological Association (APA; APA Division 12, 1993). However, to avoid dependence on expert raters and operationally define the ratings for each *SoM* scale, some previously ill-defined constructs required operationalization (e.g., the statistical mechanisms for demonstrating *Construct Validity*). Defining these constructs ended up being one of the major challenges of this project and is discussed more below.

In the end, none of the 11 *Assessment Instruments* evaluated by the *SoM* received an "*Adequate*" rating or better on each of the *SoM* scales and no instrument received an "*Adequate*" rating on either the Criterion based/Elaborative Validity or Treatment Utility scales. These results are highly concerning and unexpected given that some of the most well regarded *Assessment Instruments* used in the fields of psychology are represented (e.g., the Wechsler Intelligence Scale for Children, the Child Behavior Checklist).

When the results of the *SoM* are examined more closely, they tell a more nuanced story. First, it is noteworthy that the ratings of the lower order scales (psychometric dimensions established prior to widespread use of the *Assessment Instrument*; i.e., *Readability*, *Standardization*, and *Normative Data*) are higher than the higher order scales (dependent on the lower order scales; i.e., the two *Validity* scales and *Treatment Utility*) with the intermediate

scales in the middle (i.e., the three *Reliability* scales). In fact, all of the ratings for *Standardization* were "*Excellent*," indicating that the "procedures for administering, scoring, and interpreting scores are included on the measure or in the administration/scoring manual" and all but one *Assessment Instrument* for which readability was relevant received a "*Good*" or "*Excellent*" rating. The implication of these findings may be that the poor performance of the *Assessment Instruments* on the *SoM* is due to inadequate research rather than inherent problems with the *Assessment Instruments* themselves. If this is the case, then the development, acceptance, and use of a system for evaluating *Assessment Instruments* like the *SoM* may lead to targeted research evaluating and potentially demonstrating the strength of *Assessment Instruments* on these higher order scales.

During the development of the *SoM*, it was decided that the ratings for multi-scale *Assessment Instruments* (e.g., the Achenbach scales) would be determined by the lowest-rated scale. For example, in an *Assessment Instrument* with five scales receiving "*Good*" ratings for four of the five scales and "*Less than Adequate*" for the final scale, the overall rating for the scale would be "*Less than Adequate*." This stringent scoring system was used because the goal of the *SoM* was to provide a minimal number of data points to describe the psychometric strength of an instrument and the alternatives, reporting the ratings for each scale or reporting either the mean or median score, seemed too cumbersome and misleading respectively. This scoring rule applies to Internal Consistency, Inter-rater Reliability, Test-retest Reliability, Construct/ Representative Validity, and Criterion Related/Elaborative Validity. Further work on the *SoM* should readdress whether assigning the entire Assessment Instrument a rating based on the rating of its lowest scale is the best option.

Another critical and innovative decision that was made during the development of the *SoM* addressed the distinction between *Construct/Representative Validity* and *Criterion-related/Elaborative Validity*. As discussed in earlier in this document, construct validity essentially refers to the way in which the construct(s) measured behaves relative to other related constructs and is essentially measured via correlations (Cronbach & Meehl, 1995; Foster & Cone, 1995). For this reason, the primary statistical mechanism used by the *SoM* for demonstrating *Construct/Representative Validity* is a correlation (i.e., the presents of a correlation for convergent validity or the absence of a correlation for discriminant validity). Additionally, the use of a confirmatory factor analysis (CFA) was also considered an acceptable mechanism for demonstrating *Construct/Representative Validity* when evaluating multi-scale *Assessment Instruments* so long as the factor structure tested by the CFA matched the factor structure proposed by the *Assessment Instrument's* manual and used by its scoring system. Finally, a multitrait-multimethod matrix with predicted correlations was considered evidence for the *Construct/Representative Validity* scale.

Whereas construct validity refers primarily to correlations, criterion-related validity essentially refers to whether the instrument or scale in question accurately predicts other important events (Foster & Cone, 1995). Therefore, ratings on the *Criterion-Related/Elaborative Validity* scale of the *SoM* are determined by the presence of expected predictions of events in the literature. This can include scores on other established measures of the same or similar constructs. It was decided that the use of statistical tests, such as *t*-tests and analyses of variance, comparing the mean of two different groups would not be considered support for either type of validity. These decisions represent a first effort to clearly define which statistical tests support the various psychometric dimensions evaluated in the *SoM*. That is, correlations, factor analyses,

and multitrait-multimethod matrix supported *Construct/Representative Validity* scale and regressions supported *Criterion-Related/Elaborative Validity* scale. Future efforts to improve the *SoM* should revisit the operational definitions for each of the *SoM* scales with a particular eye toward the two validity scales, the mechanisms for demonstrating support for these two types of validity, and whether the current definitions best achieves the goals of the *SoM*.

Comparing the Result of the SoM to Other Assessment Reviews

Another way to evaluate the *SoM* as a measure is by comparing the ratings of an *Assessment Instrument* when evaluated by the *SoM* versus other reviews of the same *Assessment Instrument*. The findings of three other assessment evaluation efforts warrant consideration: (a) articles from the 2005 special section on EBA in the Journal of Clinical Child and Adolescent Psychology (JCCAP); (b) chapters from Mash and Barkley's 2007 edited book titled *Assessment of Childhood Disorders*; and (c) Hunsley and Mash's 2008 book titled *A Guide to Assessments that Work*.

The Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001), for example, is a widely used, parent-report instrument that includes three broadband scales and eight syndrome scales. For the purposes of the *SoM*, the CBCL was evaluated on the eight syndrome scales and received ratings of "*Excellent*" on the *Readability* and *Standardization* scales, "*Good*" on the *Normative Data* scale, and "*Adequate*" on the *Internal Consistency* and *Test-retest Reliability* scales. The CBCL was rated "*Less than Adequate*" on the Construct/Representative Validity and *Criterion-Related/Elaborative Validity* scale. Inter-rater Reliability was "*Not Applicable*" to the CBCL. Overall, the CBCL did not meet the "*Adequate*" level of better on each of the *SoM* scales but received the highest mean rating across the nine *SoM* scales—1.25.

The CBCL was also reviewed by each of the three sources listed above. These three sources are organized around diagnosis rather than the instrument itself. As a result, the information on the CBCL is spread across multiple articles and chapters in these sources. For example, in the JCCAP special section, the CBCL was evaluated relative to its use with depression (Klein, Dougherty, & Olino, 2005), autism spectrum disorders (Ozonoff, Goodlin-Jones, & Solomon, 2005), attention deficit/hyperactivity disorder (Pelham, Fabiano, & Massetti, 2005), anxiety disorders (Silverman, Ollendick, 2005), and bipolar disorder (Youngstrom, Findling, Youngstrom, Calabrese, 2005).

With regard to the CBCL's use for assessing depressive symptoms, Klein et al. (2005) and Dougherty and colleagues (Dougherty, Klein, Olino, & Laptook, 2008) express concern with the CBCL since it does not include a scale explicitly focused on depression and therefore is less useful for this purpose. Additional concerns about the low reliability of the depression items in the CBCL are raised by Rudolph and Lambert (2007). Perhaps because of the poor match between the CBCL scales and the construct of depression, none of the above sources make specific recommendations related to the use of the CBCL for depression. It can be said, however, that the general tone of the commentary regarding the CBCL as a measure of depression is guarded.

With regard to the CBCL as a measure of conduct related behaviors, the Achenbach System of Empirically Based Assessment (ASEBA) which includes the CBCL is recommended for use for screening purposes, assessing overt and covert conduct related problems, and assessing for comorbid problems (McMahon & Frick, 2005). When used as an instrument to support diagnosis, Frick and McMahon (2008) "*Highly Recommended*" the ASEBA (including the CBCL, Youth Self-Report [YSR], and Teacher Report Form [TRF]). The ASEBA received

rating of "*Excellent*" on "*Norms*", "*Internal Consistency*", "*Construct Validity*", and "*Validity Generalization*", a rating of "*Good*" for "*Content Validity*", and Average ratings for "*Inter-Rater Reliability*" and "*Treatment Utility*." While these scale do not directly relate to the scales used in the *SoM* and the focus of the ratings by Frick and McMahon (2008) is the use of the ASEBA used for Conduct Problems versus the CBCL as a whole, the ratings provided by Frick and McMahon are substantially higher than those provided by the *SoM*.

Comparing these results further supports the idea that the *SoM* is a strict judge of psychometric strength. However, one cannot assume that the cause of the discrepancy between the *SoM* and other assessment evaluation efforts is solely the inappropriate strictness of the *SoM*. The review articles in the JCCAP's special section (2005) and the chapters in Mash and Barkley's book (2007) did not include which dimension were being used to evaluate the *Assessment Instruments* and the criteria used to judge the strength of the *Assessment Instruments*. On the other hand, the structure of these three sources, primarily disorder based versus *Assessment Instrument* based, does have some benefits. For example, it allows users review the *Assessment Instruments* useful for a particular purpose rather than having all of the *Assessment Instruments* clustered together. Additionally, supports the review of scales within an *Assessment Instrument* rather than the *Assessment Instrument* as a whole. This would have the beneficial result of not lowering the rating for an entire *Assessment Instrument* based on one poorly performing scale. On the other hand, it could lead to clinicians assuming that an entire *Assessment Instrument* possesses the particular psychometric strengths reported for a single scale within the *Assessment Instrument*. Future work with the *SoM* should revisit the pros and cons of organizing the *SoM* by *Assessment Instrument* versus diagnosis and scales.

Strengths and Weaknesses of the Reports

As noted in the results section above, none of the five hypotheses in this study were supported. That is, none of the instruments used in the Reports were rated "*Adequate*" or better on the *SoM* scales (hypothesis one), each *Reason for Referral* was not assessed by at least two *Assessment Instruments* (hypothesis two), each Primary Diagnostic category was not assessed my at least one *Assessment Instrument* (hypothesis three), and each report did not include at least two *Informants* (hypothesis four) and *Methods of Assessment* (hypothesis five). Each of the hypotheses were based on criterions set at levels consistent with the bulk of the literature on EBA (e.g., Mash &Hunsley, 2005a; Hunsley & Mash, 2008; Cronbach and Meehl, 1955; Streiner, 2003; Sattler, 2008; Foster & Cone, 1995) and best practices in psychological assessments (e.g., Angold & Fisher, 1999; Jensen-Doss, Cook & McLeod, 2007). The failure to meet these criteria needs to be interpreted carefully and in light of the current state of the research in the field of EBA and current assessment practices in the field. It is also important to consider how far from meeting these criteria the reports actually were. Below is a discussion of the findings in terms of the *Reason for Referral*, *Primary Diagnosis*, *Informant*, and *Method of Assessment* variables.

Assessing reason for referral. As noted earlier, hypothesis two, evaluating the number of *Assessment Instruments* addressing each *Reason for Referral*, was false simply because there were no *Assessment Instruments* deemed "*Adequate*" or better on the *SoM*. Even with less-stringent criteria addressed in the post hoc analyses, removing the requirement that the *Assessment Instruments* be deemed "*Adequate*" or better on the *SoM*, not every Report included two *Assessment Instruments* addressing each *Reason for Referral*. While these results clearly reflect suboptimum assessment practices, they do not capture the positive aspects of the Reports.

Table 16 (see above) shows that 11 of the 32 Reports included three or more *Assessment Instruments* addressing the *Reason for Referral Domain* and an additional six including two *Assessment Instruments* addressing the *Reason for Referral Domain*. In all, 30 Reports were coded as having a specific *Reason for Referral* with 17 (56.67%) meeting the criterion and an additional six Reports (20.0%) including one *Assessment Instrument* addressing the *Reason for Referral Domain*. On the seven Reports that did not include any *Assessment Instruments* addressing the *Reason for Referral Domain*, five were for youth whose *Primary Reason for Referral Domain* was *Caregiver Behavior*. This leaves only two Reports which included a *Primary Reason for Referral Domain* that was youth oriented (i.e., addressing the youth's behaviors rather than a caregiver's behaviors or circumstances) that did not include any *Assessment Instruments* addressing the Reason for Referral. If this CMHC's Assessment Clinic included assessment of caregiver mental health and behavioral concerns in their assessment practice, then they could go a long way toward closing the gap between their current practice and what is considered best practice associated with addressing the *Reason for Referral*.

Assessing primary diagnosis. The data related to the Assessment Clinic's use of *Assessment Instruments* to address each of the *Primary Diagnostic Categories* are very promising. Specifically, each of the 26 Reports which included a *Primary Axis 1 Diagnosis Domain* (not including diagnoses included in the *Other* category) was assessed by at least one *Assessment Instrument*. This finding stands as a clear area of strength for the Reports.

Informants and method of assessment. The importance of using multiple sources of information and multiple methods of assessment is based on the notion that psychological problems are often multifaceted thus requiring multiple assessment instruments to capture the full scope of the problem (Kazdin, 2003; De Los Reyes & Kazdin, 2005; Jensen-Doss et al.,

2007). Additionally, different informants are differentially able to identify, and therefore, report on different aspects of an individual's experience. Finally, individuals respond differently to different methods of assessment (e.g., self-report versus interview; Kazdin, 2003).

In reviewing the data from the 32 Reports used for this study, it was revealed that, while the criterion of multiple *Informants* and multiple *Methods of Assessment* for each Report was not met, the majority of the Reports did meet this standard. Specifically, 71.88% of the Reports included two or more *Informants* when only *Identifiable/High Use* and *Identifiable/Low Use Assessment Instruments* were considered. This analysis did not consider *Assessment Components* that were *Unidentifiable/Specific Practice* and or considered *General Practice* assessment because the psychometric characteristics of the *Unidentifiable/Specific Practice* and *General Practice Assessment Components* are unknown. However, the decision to ignore these *Assessment Components* does result in an underestimate of the number of informants providing information to the assessors. Additionally, since data were not available to determine which parent or teacher was the informant or if multiple parents or multiple teachers contributed to the assessment, it is possible that multiple *Informants* from the same category (e.g., two parents or two teachers) provided assessment data but that this was not captured. All in all, the majority of the Reports met the standard set forth by hypothesis four even though the Reports collectively did not.

The data on the number of *Methods of Assessment* used in each of the Assessment Clinic's Reports are even more promising. Like the analysis of number of *Informants* per Report, the analysis addressing the number of *Method of Assessment* did not consider *Unidentifiable/Specific Practice* or *General Practice Assessment Components* for the analysis for the same reason as noted in the previous paragraph. In this case, 93.75% of the Reports included

more than one *Method of Assessment*. None of the *Assessment Instruments* identified in the Reports were coded as a "*Clinical Interview*." Thus, if the "*Youth Interviews*" or "*Meetings*", frequently listed as *Assessment Components*, were included in the analysis, then the number of *Methods of Assessment* would have been even higher. While the standard of two or more *Methods of Assessment* per Report was not met, this remains an area of strength on the part of the CMHC's Assessment Clinic.

It is also worth noting that the here that the decision to not include "*Youth Interviews*" was made solely on the basis that there are not data on the psychometric characteristics of this assessment strategy rather than an blanket rebuke of non-standardized assessment interviews. Unstructured assessment interviews provide important information to clinicians (Jensen-Doss et al., 2007) and it is unlikely that their use will disappear anytime soon. Among the advantages of unstructured interviews is the flexibility afforded to clinicians so that they can customize the questions for the individual and follow-up on topics (e.g., such as antecedents to problem behaviors and other environmental factors) in a way not possible with standardized interviews (Jensen-Doss et al., 2007). On the other hand, it is exactly this flexibility and clinician-to-clinician variability that makes reliance on unstructured interviews problematic.

Limitations

The present study has contributed to the EBA effort through the development and piloting of criteria for evaluating *Assessment Instruments* and an initial effort to evaluate the assessment practices in a CMHC's Assessment Clinic. Despite these contributions, there are several limitations to consider as well. First, the sample used for this study consisted of the first 32 youth who received services from a newly opened Assessment Clinic at a CMHC over the course on one year. While this sample likely is representative of the population of youth

receiving assessment services at this CMHC, it is insufficient for three primary reasons. First, the sample is too small to provide the needed power for many statistical tests. Second, the sample is unlikely to be representative of the larger population of youth receiving assessment services beyond this particular CMHC. Third, the Reports represent the work of only two clinical psychologists working in conjunction with one social worker and therefore the services provided likely represent the practices of these three individuals and not the larger community of providers of assessment services. Additionally, since both of the clinical psychologists working for the Assessment Clinic were recent graduates, there is the possibility of a cohort effect. That is, it is conceivable that early career psychologists engage in different practices that those who have been in clinical practice for a longer period of time. While there is evidence suggesting that therapist's experience with psychotherapy does not predict outcome (Weisz, Weiss, Han, Granger, & Morton, 1995), the influence of therapist's experience on assessment services is unknown. Finally, only 11 *Assessment Instruments* were evaluated with the *SoM* in this study. This limited sample of *Assessment Instruments* limited the conclusions in two major respects. First, with such a small sample size, it is difficult to determine if the results to the *SoM* on these 11 *Assessment Instruments* are representative of the psychological Assessment Instruments in general. Second, the small sample size limited the ability of statistical tests to detect real relationships (e.g., the correlation between *SoM* rating and frequency of use).

Due to the small sample size used for this study impacted the statistical tests that could be run. Specifically, the small sample size and limited number of *Assessment Instruments* rated with the *SoM* impacted the correlation between *SoM* scores and frequency of use and all of the *t*-tests used in the post hoc analysis of the *Reason for Referral* and *Primary Diagnosis*.

A second limitation to note with this study concerns the search for journal articles addressing the *Assessment Instruments* reviewed by the *SoM*. The ratings on the *SoM* scales, and hypothesis one, two, and three in turn, hinged heavily on the identification of relevant literature documenting the psychometric strength of the assessment instruments used. A number of methodological decisions were made related to the literature review for this study that likely impacted the results. First, only articles published since 1995 were included. This decision was made because of the changes in demographic characteristics of the population, changes in social norms, the Flynn Effect, and the publication of the Fourth Edition of the Diagnostic and Statistical Manual all could diminish the validity of older data. However, measures that were developed, normed, and primarily studied prior to this date were penalized by this decision. Additionally, newly developed and published measures that have not had a chance to accrue published studies evaluating their psychometric strength were also penalized. A number of well regarded measures, such as the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV) and Children's Depression Inventory (CDI), were evaluated by surprisingly few recent studies. The WISC-IV and the CDI are the most widely used *Assessment Instruments* addressing intelligence and depressive symptoms respectively (Ozonoff et al., 2005; Klein et al., 2005). This frequency of use may have led to overconfidence in these *Assessment Instruments* and impede additional research on their psychometric strength. The lack of research evaluating these *Assessment Instruments* had the effect of limiting their ratings on the *SoM* scales that required independent data (e.g., the validity scales). Finally, the lack of support for *Treatment Utility* may be the result of not explicitly including the term "utility" in the search field.

There are several limitations related to the *SoM* scale. First, it is important to remember that the *SoM* measure the psychometric strength of an *Assessment Instrument* and not how the

Assessment Instrument was used. Therefore, the *SoM* scales allow for an evaluation of the *Assessment Instruments* used to complete an assessment report but cannot evaluate (1) the fidelity to the standardized procedures with which the instruments were administered or (2) accuracy with which the instruments were interpreted.

Another limitation of the *SoM* scales was the low rating that each of the instruments received on the scales and the relative lack of variability of scores. On the one hand, the *SoM* was developed as a criterion-based scale, meaning that the different ratings correspond to bench marks of increasing psychometric strength. Thus, the measure was not designed to have normally distributed item scores. These benchmarks were created based on standards recommended in the existing literature on psychometrics (e.g., Mash & Hunsley, 2005a; Hunsley & Mash, 2008; Cronbach & Meehl, 1955; Streiner, 2003; Sattler, 2008; Foster & Cone, 1995). Thus, the failure to meet these standards is a shortcoming of either the instrument characteristics or the research on the instrument and compensations should not be made for these shortcomings. On the other hand, the "*Less than Adequate*" ratings assigned to well regarded measures may be an indication that either the criteria are too strict or that the literature review failed to identify important support for these measures. Either way, the lack of variability of ratings on the *SoM* resulted in a failure of the measure to differentiate instruments. For example, the ratings indicate that there are no meaningful differences among the instruments on *Criterion-Related/Elaborative Validity* or *Treatment Utility*.

Finally, the low rating on the *SoM* by the *Assessment Instruments* undermined the testing of hypotheses two through five since they were dependent on the Assessment Instrument being deemed "*Adequate*" or better on the *SoM*.

Future Directions

Despite the limitations of this study and the preliminary nature of the *SoM*, this study has opened up a rich area of research related to evidence-based assessment (EBA). Studies to follow certainly need to address the mechanical problems related to sample size and limited number *Assessment Instruments* evaluated with the *SoM*. Beyond those issues, ongoing refinement of the *SoM* is the most critical issue facing researchers attempting to take a systematic approach to evaluate *Assessment Instruments* against clearly defined and empirically-supported, criterion-based standards. Finally, improvements to the *Reason for Referral* and *Primary Diagnosis Category* variables and coding procedures will improve the confidence with which conclusions are made regarding the selection of *Assessment Instruments* for youth with various problems. These three issues are discussed briefly.

Future studies using the *SoM* or those evaluating assessment practices in clinical practice more generally should work toward improving the sample in at least two respects. First, a larger sample which includes more youth Reports would increase the power of statistical tests to detect relationships and differences. Second, the sample should include Reports from a variety of service centers such as CMHC, local and regional hospitals, psychology training clinics, and private practice. Increasing the size of the sample in either of these ways would have the effect of increasing the number of clinicians providing the assessment services. All of these steps would improve the field's understanding of what might be considered usual care assessment in each of these settings.

Additional work is needed to calibrate the *SoM* scales with an eye toward maintaining its function as a criterion-based measure while improving its ability to differentiate between the quality of different instruments. The *SoM* scales were identified and the criteria for the different

levels of support (i.e., "*Less than Adequate*", "*Adequate*", "*Good*", and "*Excellent*") based largely from review of the psychometric literature and past efforts to create assessment evaluation procedures (e.g., Mash & Hunsley, 2005a; Hunsley & Mash, 2008; Cronbach & Meehl, 1955; Streiner, 2003; Sattler, 2008, Foster & Cone, 1995). Future modifications to the *SoM* should consider the development of a wider range of scores that then map onto the criterion levels used in the present form of the *SoM*. This would provide the added advantage of making the measure useful when no measure is available that meets the "*Adequate*" level of support. Additionally, this type of scoring system would allow for the criterion cut off scores to be empirically determined as the field continues to develop an understanding of what measures have "*Adequate*" *Treatment Utility*. Ongoing work with the *SoM* should also consider whether the current dimensions are the most appropriate. For example, it may be unnecessary for the *SoM* to include three scales addressing reliability and two scales addressing validity. This is especially true of the two validity scales since the statistical means of differentiating *Construct/ Representative Validity* from *Criterion-Related/Elaborative Validity* are so similar.

For this study, youth's *Reason for Referral* was coded based on information provided on the IRFs and the portions of the Reports. This coding was important to the second hypothesis, that each youth would receive at least two *Assessment Instruments* addressing the youth's *Reason for Referral*. This line of research would benefit from the collection of *Reason for Referral* information prior to the initiation of the assessment to ensure that the *Reason for Referral* code is not influenced by data collected after the start of the assessment. An alternative method of identifying Reason for Referral used in previous research involved collecting data directly from the youth being assessed and the youth's parent and categorizing it into codes corresponding to CBCL items (Yeh & Weisz, 2001). This procedure, while supported by

precedent, still would require the organization of codes into categories corresponding to domains assessed by *Assessment Instruments*.

Conclusion

In the early 1990's, Division 12 of APA published a set of guidelines establishing the criteria for evaluating the efficacy of treatments. That move sparked a substantial move toward developing, researching, disseminating, and ultimately, implementing evidence-based psychological treatments in communities around the country. However, the EBT criteria stated the need for reliable and valid assessment of participants in treatment studies but to not clarify what exactly constitutes sufficient reliability of validity (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). Mash and Hunsley (2005a) humorously note that this is "tantamount to writing a recipe for baking hippopotamus cookies that begins with the instruction 'use one hippopotamus,' without directions for securing the main ingredient" (Mash & Hunsley, 2005a). This study is meant to be a preliminary step down the long path toward developing a science-based system for evaluating *Assessment Instruments* and *Assessment Practices*. Hopefully assessment research will see a surge in interest and investment like what was seen in the treatment area since the Division 12 first published guidelines for establishing EBTs.

Given the overall lack of support found for the *Assessment Instruments* and general practices found in this study, the field's push toward evidence-based practice is in its infancy. This study indicates that a great deal of work is needed to improve the state of evidence-based assessment. With hard work by many, this line of research can have the same effect on psychological assessments as the Division 12 evidence-based treatment guidelines did almost 20 years ago. If nothing else, by establishing criteria by which *Assessment Instruments* are

measured, researchers will have an understanding of what studies are needed to demonstrate an *Assessment Instrument's* psychometric strength.

This line of research will also have implications beyond psychological assessments. The progress that has been made on the EBT front over the past two decades has centered on the development and identification of treatments that effectively treat specific diagnoses or problem areas. On one hand, these treatments are beginning to demonstrate superior efficacy over treatment approaches that are not diagnosis or problem specific (Weisz, Jensen-Doss, Hawley, 2005). On the other hand, the effectiveness of these treatments depends on accurate diagnoses (Jensen-Doss et al., 2007; Weisz & Addis, 2006). Therefore, improvements in the field of assessment will have a meaningful impact for youth and adults seeking treatment for mental health diagnoses and problems.

List of References

List of References

Achenbach, T.M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 541-547.

Achenbach, T.M., & Rescorla, L.A. (2001) Manual for the ASEBA school age forms and profiles. Burlington: University of Vermont, Research Center for Children, Youth, and Families.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Anastasi, A., (1988). Psychological Testing (6th ed.). New York: MacMillan Publishing Company.

American Psychological Association. (1993). Task force on promotion and dissemination of psychological procedures. [online]. Available: http://www.apa.org/divisions/div12/est/chamble2.pdf. (November 24, 2010)

American Psychological Association. (2005a). Policy Statement on Evidence-Based Practice in Psychology. [online]. Available: http://www.apa.org/practice/resources/evidence/evidence-based-statement.pdf. (November 24, 2010)

American Psychological Association. (2005b). Report of the 2005 presidential task force on evidence-based practice. [online]. Available: http://www.apa.org/practice/resources/evidence/evidence-based-report.pdf. (November 24, 2010)

Angold, A., & Fisher, P. W. (1999). Interviewer-based interviews. In Shaffer, D., Lucas, C. P., & Richters, J. E. (Eds), *Diagnostic Assessment in Child and Adolescent Psychopathology* (pp. 34-64). New York, NY: The Guilford Press.

Barkley, R. A., (1997). *Defiant children: A clinician's manual for assessment and parent training* (2nd ed.). New York, NY: Guilford Press.

Barry, C. T., & Pickard, J. D. (2008). Developmental Issues. In Hersen, M., & Reitman, D. (Eds.), *Handbook of Assessment, Case Conceptualization, and Treatment Volume II: Children and Adolescents* (pp. 76-101). Hoboken, NJ: John Wiley & Sons.

Bridges, A.J., & Holler, K.A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology, 13*, 528-538.

Chambless, D.L., & Hollon, S.D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7-18.

Chambless, D.L., & Ollendick, T.H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review in Psychology, 52*, 685-716.

Cicchetti, D.V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551-558.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, New Series, 243*(4899), 1668-1674.

DeVellis, R.F., (2003) *Scale development: Theory and application* (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Dougherty, L.R., Klein, D.N., Olino, .T.M, & Laptook, R.S. (2008). *Depression in Children and adolescents*. In J. Hunsley & E.J. Mash (Eds) A guide to assessments that work. New York: Oxford University Press.

Fleiss, J. L., (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378-382.

Foster, S. F., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248-260.

Freeman, K.A., & Miller, C.A. (2002). *Behavioral case conceptualization for children and adolescents*. In M. Hersen (Eds.), *Clinical Behavior Therapy: Adults and Children* (pp. 239-255). New York: Wiley.

Frick, P.J., & McMahon, R.J. (2008). *Child and adolescents conduct problems*. In J. Hunsley & E.J. Mash (Eds) A guide to assessments that work. New York: Oxford University Press.

Groth-Marnat, G., (2003). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley & Sons Inc.

Hayes, S.C., Nelson, R.O., & Jarrett, R.B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42*(11), 963-974.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Holmbeck, G. N., Greenley, R. N., & Franks, E. A. (2003). Developmental issues and considerations in research and practice. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children & adolescents* (pp. 21-40). New York: Guilford.

Hunsley, J. & Mash, E. J., (2008). *A guide to assessments that work*. New York, NY: Oxford University Press.

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, *70*(1), 158-168. doi: 10.1037//0022-006X.70.1.158.

Doss, A. J., Cook, K. T., & McLeod, B. D. (2008). Diagnostic issues. In Hersen, M., & Reitman, D. (Eds.), *Handbook of Assessment, Case Conceptualization, and Treatment Volume II: Children and Adolescents* (pp. 25-52). John Wiley & Sons.

Kazdin, A.E. (2003). *Research design in clinical psychology*, 4th edition. Boston, MA: Allyn & Bacon.

Kazdin, A.E. (2007). *Assessment and Evaluation in Clinical Practice*. In C. D. Goodheart, A. E. Kazdin, R. J. Sternberg (Eds) Evidence-based psychotherapy: Where practice and research meet (pp. 179-206). Washington, DC: American Psychological Association.

Kendall, P. C. & Norton-Ford, J. D., (1982). *Clinical psychology: Scientific and professional dimensions*. New York, NY: Wiley.

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin, 65,* 110-136.

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of clinical child and adolescent psychology : the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *34*(3), 412-32. doi:10.1207/s15374424jccp3403_3

Lantz, C. A. & Nebenzahl E. (1996). Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*. 49(4):431-434.

Levant, R. F., Barlow, D. H., David-, K. W., Hagglund, K. J., Hollon, S. D., Johnson, J. D., et al. (2006). Evidence-based practice in psychology. *The American Psychologist*, *61*(4), 271-285. doi: 10.1037/0003-066X.61.4.271.

Mash, E. J., & Barkley, R. A. (Eds) (2007). *Assessment of childhood disorders (4th Ed.)*. New York, NY US: Guilford Press.

Mash, E. J. & Hunsley, J. (2005a). Evidence-based assessment of child and adolescent disorders: Issues and challenges. Journal of Clinical Child and Adolescent Psychology, 34(3), 362-379.

Mash, E. J. & Hunsley, J. (2005b). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. Psychological Assessment, 17(3), 251-255.

Mash, E. J., Hunsley, J. (2007). Assessment of child and family disturbance: A developmental-systems approach. In Mash, E. J., & Barkley, R. A. (Eds) *Assessment of childhood disorders (4th Ed.)* (pp. 3-50). New York, NY US: Guilford Press.

McHugh, R. K. & Behar, E. (2009) Readability of Self-Report Measures of Depression and Anxiety. *Journal of Consulting and Clinical Psychology, 77*(6), 1100-1112.

McMahon, R.J., & Frick, P.J., (2007). Conduct and oppositional disorders. In Mash, E. J., & Barkley, R. A. (Eds) *Assessment of childhood disorders (4th Ed.)* (pp. 132-183). New York, NY US: Guilford Press.

Meehl, P.E. (1956). Symposium on clinical and statistical prediction: The tie that binds. *Journal of Counseling Psychology*, 3(3), 163-164.

Miller, A. L., Rathus, J. H., Linehan, M. M., (2007). *Dialectic behavior with suicidal adolescents*. New York, NY: Guilford Press.

Nelson-Gray, R.O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15*(4), 521-531.

Pelham, W. E., Fabiano, G. a, & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 449-76. doi:10.1207/s15374424jccp3403_5

Persons, J. B., (1989). *Cognitive therapy in practice: A case formulation approach*. New York, NY: W W Norton & Co.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 523-40. doi:10.1207/s15374424jccp3403_8

Robinson, J. P., Shaver, P. R., Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In J. P. Robinson, P. R. Shaver and L. S. Wrightsman (Eds.) *Measures of personality and social psychological attitudes* (1-16). San Diego, CA: Academic Press.

Rudolph, K.D., & Lambert, S.F. (2007). Child and adolescent depression. In Mash, E. J., & Barkley, R. A. (Eds) *Assessment of childhood disorders (4th Ed.)* (pp. 231-252). New York, NY US: Guilford Press.

Rutter, M. & Sroufe, L.A. (2000). Development and psychopathology: Concepts and challenges. *Development and Psychopathology, 12*, 265-296

Sattler, J.M., & Hoge, R.D. (2006). *Assessment of children: Behavioral, social, and clinical foundations* (5th ed.). La Mesa, California: Jerome M. Sattler Publisher, Inc.

Sattler, J.M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). La Mesa, California: Jerome M. Sattler Publisher, Inc.

Sattler, J.M., & Hoge, R.D. (2006). *Assessment of children: Behavioral, social, and clinical foundations, 5th edition*. La Mesa, California: Jerome M. Sattler Publisher, Inc.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Shrout, P.E., Spitzer, R.L., & Fleiss, J.L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44*, 172-177.

Shaffer, D., Fisher, P. W., & Lucas, C. P. (1999). Respondent-based interviews. In Shaffer, D., Lucas, C. P., & Richters, J. E. (Eds), *Diagnostic Assessment in Child and Adolescent Psychopathology* (pp. 3-33). The Guilford Press.

Silverman, W. K., & Ollendick, T. H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 380-411. doi:10.1207/s15374424jccp3403_2

Smith, G.T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396-408.

Smith, R. (1991). Where is the wisdom…? The poverty of medical evidence. *British Medical Journal, 303*, 798-799.

Sommers-Flannagan, J., & Campbell, D.G. (2009). Psychotherapy and (or) medications for depression in youth? An evidence-based review with recommendations for treatment. *Journal of Contemporary Psychotherapy, 39*, 111-120.

Streiner, D.L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103.

Uebersax, J.S. (1987) Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1), 140-146.

Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

Weisz, J.R., & Addis, M.E. (2006). The research-practice tango and other choreographic challenges: Using and testing evidence-based psychotherapies in clinical care settings. In C. D. Goodheart, A. E. Kazdin, R. J. Sternberg (Eds) Evidence-based psychotherapy: Where practice and research meet (pp. 179-206). Washington, DC: American Psychological Association.

Weisz, J.R., Doss, A.J., & Hawley, K.M. (2005). Youth psychotherapy outcome research: A Review and critique of the evidence base. *Annual Review of Psychology*, *56*, 337-363. doi: 10.1146/annurev.psych.55.090902.141449.

Weisz, J.R., Jensen-Doss, A., & Hawley, K.M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. American Psychologist, 61, 671-689.

Weisz, J.R., Bahr, W., Han, S.S., Granger, D.A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117*(3), 450-468. doi: 10.1037/0033-2909.117.3.450

Yeh, M., & Weisz, J R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology*, *69*(6), 1018-1025. doi:10.1037//0022-006X.69.6.1018

Youngstrom, E.A, Findling, R.L., Youngstrom, J.K., & Calabrese, J. R. (2005). Toward an evidence-based assessment of pediatric bipolar disorder. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 433-48. doi:10.1207/s15374424jccp3403_4

Appendix A

Table 18

Strength of Measure Variable: Readability

Strength of Measure Variable: Readability	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 19

Strength of Measure Variable: Standardization

Strength of Measure Variable: Standardization	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 20

Strength of Measure Variable: Normative Data

Strength of Measure Variable: Normative Data	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 21

Strength of Measure Variable: Internal Consistency

Strength of Measure Variable: Internal Consistency	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 22

Strength of Measure Variable: Inter-rater Reliability

Strength of Measure Variable: Inter-rater Reliability	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 23

Strength of Measure Variable: Test-retest Reliability

Strength of Measure Variable: Test-retest Reliability	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 24

Strength of Measure Variable: Construct/Representative Validity

Strength of Measure Variable: Construct/ Representative Validity	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 25

Strength of Measure Variable: Criterion-related/Elaborative Validity

Strength of Measure Variable: Criterion-related/Elaborative Validity	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Table 26

Strength of Measure Variable: Treatment Utility

Strength of Measure Variable: Treatment Utility	**Definition**
Not applicable	**This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.**
Less than Adequate	
Adequate	
Good	
Excellent	

Appendix B

SoM Coding Sheet

Coder:_____ **Assessment Instrument:** _____
Publication Date: _____

Author 1: _ _____
Author 2: _ _____
Author 3: _ _____
Author 4: _ _____

Scope of this evaluation (what scales and/or scoring system is being evaluated):
Scoring system: _____

Scales (include very brief descriptor if scale name is not face valid):
1: _____
2: _____
3: _____
4: _____
5: _____
6: _____
7: _____
8: _____
9: _____
10: _____
11: _____
12: _____
13: _____
14: _____
15: _____

Factor structure proposed in the manual:

Note: Citations, for articles, on this coding sheet should include: first author, publication date, and page number(s). (E.g., Veltri, 2009, pp. 293-294)

Citations, for the manual, should include Manual, first author, publication date, and page number(s). (E.g., Manual, Butcher, 1992, pp. 27-30)

Informant

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Method of Assessment

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Readability

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Standardization

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Normative Data

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Internal Consistency

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Inter-rater Reliability

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Test-retest Reliability

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Construct/Representative Validity

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Criterion-related/Elaborative Validity

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Treatment Utility

This portion of the document was removed for this publication. Please contact the author Cassidy C. Arnold (Cassidy.C.Arnold@Gmail.com) with questions or for additional information.

Appendix C

Reason for Referral Coding Sheet

ID: _____ Date: _____ Coder: _____

The Assessment Clinic was developed to provide assessment and diagnostic services to youth who are at risk of out-of-home placements (e.g., foster care or residential treatment centers). We are interested in determining why the youth was originally referred for an assessment at the Assessment Clinic. Please make an effort to code only the information describing the reason why the youth is being referred to the Assessment Clinic that is presented in the Initial Referral Form and the Assessment Clinic Report and do not interpret beyond what is explicitly stated. That is, what problem or problems are raising the possibility of an out-of-home placement? This code sheet is **not** meant to identify every problem that the youth or family is experiencing and is **not** meant to capture the findings of Assessment Clinic or outcome diagnoses.

On the reverse side of this page, you will find 12 potential reasons for referral. Please review the Initial Referral Form (not always available) and the Assessment Clinic Report to identify the reason or reasons for referral to the Assessment Clinic. Please indicate the primary reason for referral by placing a "1" in the rating column on the left side of the table on the reverse side of this form. For youth that have multiple reasons for referral, subsequent reasons for referral can be indicated with a "2" and "3".

Ratings	Reason for Referral	Description and Characteristic Diagnoses and Behaviors
	Youth's mood related symptoms	Referral to the Assessment Clinic due to the youth's depressive and/or manic symptoms. Characteristic diagnoses and behaviors include: Major Depressive Disorder, Bipolar, and Adjustment disorder with disturbance in mood.
	Youth's attention/ impulsive problem behaviors	Referral to the Assessment Clinic due to the youth's symptoms of inattention, impulsivity, or hyperactivity. Characteristic diagnoses and behaviors include: Attention Deficit/Hyperactivity Disorder.
	Youth's anxiety problems	Referral to the Assessment Clinic due to the youth's symptoms of anxiety. Characteristic diagnoses and behaviors include: Various anxiety disorder and disorders conceptually related to anxiety such as Trichotillomania, Impulse-Control Disorder, and eating disorders.
	Youth's psychotic or thought disordered symptoms	Referral to the Assessment Clinic due to the youth's psychotic or thought disordered symptoms. Characteristic diagnoses and behaviors include: Schizophrenia, Delusional Disorder, and psychotic behaviors (regardless of diagnosis)
	Youth's reckless or dangerous behaviors	Referral to the Assessment Clinic due to the youth's behaviors that risks their own health and safety Characteristic diagnoses and behaviors include: suicidal, self-injurious behavior, and/or risky behavior that puts the youth at risk of harm (e.g., reckless sexual behavior). Note: do not code substance use here—see below.
	Youth's substance use	Referral to the Assessment Clinic due to the youth's substance use and/or behaviors associated with substance use.
	Youth's conduct related behavior	Referral to the Assessment Clinic due to the youth's dangerous or destructive behaviors. Characteristic diagnoses and behaviors include: Conduct Disorder, Oppositional Defiant Disorder, fire setting, aggression, running away, and/or destructive behaviors.
	Youth's impairment	Referral to the Assessment Clinic due to the youth's cognitive, language, developmental or physical impairment.
	Caregiver's behaviors	Referral to the Assessment Clinic due to the caregiver's behaviors or characteristics. This includes abuse and/or abandonment (i.e. active caregiver behavior that harms the youth) or neglect, caregiver psychopathology, caregiver impairment, or caregiver inability (i.e., caregiver inaction that leads to insufficient care provided).
	Caregiver incarceration	Referral to the Assessment Clinic due to the caregiver being incarcerated.
	Other	Referral to the Assessment Clinic due to reasons that do not fit into any other categories.

Vita

CASSIDY C. ARNOLD

Curriculum Vitae
Virginia Commonwealth University
806 W. Franklin Street, Richmond, VA 23284
(303) 358-7072 (cell) • **arnoldcc@vcu.edu**

EDUCATION

2008 - Present **Virginia Commonwealth University**, Richmond, Virginia
Doctor of Philosophy, Clinical Psychology, anticipated Spring 2014
Master of Science, Clinical Psychology, anticipated Fall 2011
Child Track Specialization
Master's Thesis: *Evaluation of Assessment and Diagnostic Services at a Community Service Board: Development of a Novel Assessment Evaluation Tool*
Adviser: Michael Southam-Gerow, Ph.D.

2004 - 2006 **University of Colorado at Denver**, Denver, Colorado
Bachelor of Arts: Psychology

1998- 2000 **Lafayette College**, Easton, Pennsylvania
Major: Civil Engineering

CLINICAL EXPERIENCE

August 2011 – Present **Clinical Practicum Placement: Virginia Treatment Center for Children**
Virginia Commonwealth University Medical Center
Supervisor: Jaee Bodas, Ph.D.
Performed intake interviews with youth and families and conducted individual and family therapy for youth on an Inpatient Psychiatric Unit.

August 2010 – April2011 **Clinical Practicum Placement: Community Private Practice**
Glen Forest Associates
Supervisor: Creighton Hite, Psy.D., M.Ed.
Performed intake interviews, diagnostic and cognitive assessments, and wrote integrative assessment reports for child, adolescent, and adult clients.

May 2009 – Present **Clinical Practicum Placement: Anxiety Specialty Clinic**
Virginia Commonwealth University, Richmond, Virginia
Supervisors: Michael Southam-Gerow, Ph.D. & Scott Vrana, Ph.D.
2009-2010: Associate member
2010-2011: Core member

2011-Present: Chief member

Performed intake assessments and evidence-based cognitive-behavioral therapy for child, adolescent, and adult clients with Anxiety Disorders. Supervision consisted of weekly group and individual meetings. As a Chief member, provided individual and group supervision to junior Anxiety Clinic members on a weekly basis.

Jan. 2010 – May 2010 **Clinical Practicum Placement: General Clinic**
Virginia Commonwealth University, Richmond, Virginia
Supervisor: Christine Nelson, Ph.D.

Performed intake assessments and evidence-based therapy for adult clients with a variety of psychological disorders. Supervision consisted of weekly group and individual meetings.

August 2009 – Dec. 2009 **Clinical Practicum Placement: General Clinic**
Virginia Commonwealth University, Richmond, Virginia
Supervisor: Steve Auerbach, Ph.D.

Performed intake assessments and evidence-based therapy for adult clients with a variety of psychological disorders. Supervision consisted of weekly group meetings.

August 2008 – May 2009 **Clinical Didactics and Supervision**
Virginia Commonwealth University, Richmond, Virginia
Supervisor: Bryce McLeod, Ph.D.

October 2006 – July 2008 **Pediatric Mental Health Specialist**
Inpatient Psychiatric Unit, Seattle Children's Hospital
Seattle, Washington
Supervisor: Ann Moore, M.S., R.N.

Provide age appropriate coaching and supervision to patients and families while hospitalized on the Inpatient Psychiatric Unit. Implemented non-punitive interventions in an effort to support our seclusion/restraint free environment. Conducted psychosocial group sessions for child and adolescent patients with a wide range of psychological and behavioral problems along with their families.

Jan. 2006 – May 2006 **Pre-baccalaureate Practicum**
Denver Children's Home, Denver, Colorado
Supervisor: Daniela Abbott, L.M.F.T.

Co-facilitated group therapy sessions for children and adolescents at Denver Children's Home, a residential treatment center for youth who were victims of trauma, abuse and neglect.

June 2001 – August 2003 **Lead Instructor**
Outward Bound: Florida, Maine, New York
Supervisor: Pat Bunch (Florida), Wendy Jordan (Maine), Robert Burke (New York)

Facilitated and led wilderness therapy courses for at-risk and adjudicated adolescents. Taught psychosocial group sessions on topics such as anger management, problem solving, assertive communication, conflict

resolution, and self-awareness. Responsible for staff development of assistants and interns.

Summers 2004, 2005 **Lead Instructor and Course Director**
Outward Bound: Maryland
Supervisor: Cindy Ekinson and Lana Hill
Facilitated and led wilderness therapy courses for adolescents interested in developing leadership skills. Taught psychosocial group sessions on topics such as problem solving, assertive communication, conflict resolution, self-awareness, and team building skills. As course director, acted as a liaison between staff working directly with the youth, logistical staff, and families. Responsible for staff development activities for course instructors.

RESEARCH EXPERIENCE

August 2010 - Present **Research Assistant**
Treatment Integrity Measurement Study
Virginia Commonwealth University, Department of Psychology
PIs: Michael Southam-Gerow, Ph.D., Bryce McLeod, Ph.D.
The Treatment Integrity Measurement Study involves the design and development and testing of four related but distinct observational measures of treatment integrity.
As part of this project, I assisted in the development of the Adherence to Cognitive Behavioral Therapy for Youth Anxiety measure and coded therapy tapes. Additionally, I was responsible for data management on the project.

August 2008-October 2011 **Master's Thesis**: Defended October, 2011
Evaluation of a Community Mental Health Center's Assessment Clinic: Development of a Novel Assessment Evaluation Tool
This project included two components: (1), the development of a standardized, multi-dimensional Assessment Instrument evaluation tool—the Strength of Measure (SoM)—based on operationally defined criteria supported by decades of psychometric research and, (2) an evaluation of an Assessment Clinic's practices using the SoM and other criteria considered consistent with best practices.

May 2009 - August 2010 **Research Assistant**
Partnership for Excellence in Early Language and Literacy Skills (PEELLS)
Virginia Commonwealth University, Department of Education
PI: Christopher Chin, Ph.D.
The PEELLS study is a classroom-level intervention focused on promoting early language and literacy skills in Head-Start and state-funded preschool classrooms.
Conducted recruitment, consent, and early language and literacy assessments for preschool participants in classrooms in Richmond City schools, as well as data entry and analysis.

May 2009 – Jan. 2010 **Science Writer**
Commonwealth of Virginia Commission on Youth
Supervisor: Michael Southam-Gerow, Ph.D.
Summarized state of the science findings on the diagnosis, assessment, and treatment of Suicide, Non-suicidal Self-Injurious Behavior, Depression, and Trauma and Antidepressants and the Risk of Suicidal Behavior for the Commission on Youth's Collection of Evidence-Based Treatment Modalities for Children and Adolescents with Mental Health Needs, 4th Edition.

August 2008 - Present **Research Assistant**
Program Evaluation of Assessment and Diagnostic Service at Region Ten Community Service Board
Virginia Commonwealth University, Department of Psychology
PIs: Michael Southam-Gerow, Ph.D., Bryce McLeod, Ph.D.
The Program Evaluation of Assessment and Diagnostic Service at Region Ten Community Service Board was an evaluation of the assessment services with a focus on the quality of the services provided to youth at risk of out of home placement (i.e., foster care).
Developed a coding system and procedure to systematically evaluate the quality of assessments and the psychometric strength of the measures used as part of these assessments.

August 2009 - Present **Clinical Interviewer**
Adaptation of Depression and Anxiety Psychological Treatments for Children (ADAPT) Project
Community Services Board, Chesterfield County, Virginia
Supervisor: Michael Southam-Gerow, Ph.D.
The ADAPT project is an open trial of a multi-focus modular therapy for childhood with comorbid internalizing and externalizing disorders in an area community mental health clinic.
Conducted structured clinical interviews (K-SADS-PL) to evaluate youth who participated in a treatment efficacy study.

Sept. 2007 – June 2008 **Research Assistant**
Behavioral Activation Therapy for Depressed Adolescents
Seattle Children's Hospital, Department of Psychiatry and Behavioral Sciences
PIs: Elizabeth McCauley, Ph.D., Kelly Schloredt, Ph.D.
The Behavioral Activation Therapy for Depressed Adolescents study is an initial efficacy trial of manualized Behavioral Activation (BA) Therapy for depressed adolescents.
Prepared materials for the assessment component of the study.

Sept. 2007 – October 2007 **Research Assistant**
High School Transitions Study
University of Washington
PI: Elizabeth McCauley, Ph.D.
The High School Transition Study is an evaluation of a preventative intervention program focused on reducing depression, school problems,

and substance use among at-risk eighth graders entering into high school.
Supported data management on the project including data entry and data cleaning.

Feb. 2007 – June 2008 **Research Assistant**
Mental Health & Substance Use Screening Study for Teens
University of Washington, Department of Psychiatry and Behavioral Sciences
PIs: Kate Comtois, Ph.D., Elizabeth McCauley, Ph.D., Ray Hsiao, M.D.
The Mental Health & Substance Use Screening Study for Teens was an assessment validation study of the Global Appraisal of Individual Needs – Short Screener (GAIN-SS) for adolescents.
Conducted recruitment, consent, and assessments with youth participants in addition to data entry and analysis. Youth were receiving services through either inpatient or outpatient psychiatric settings.

Dec. 2006 – June 2008 **Research Assistant**
Inpatient Psychiatric Unit Eating Disorder Follow-up Study
Seattle Children's Hospital, Department of Psychiatry and Behavioral Sciences
PI: Rose Calderon, Ph.D.
The Inpatient Psychiatric Unit Eating Disorder Follow-up Study collected the health status and perspective of individuals who had formerly received services for an eating disorder in the Inpatient Psychiatric Unit at Seattle Children's Hospital.
Supported data management on the project including data entry and data cleaning.

PRESENTATIONS

Arnold, C. C. & Southam-Gerow, M. A. (2011, November). *How evidence based is usual care assessment?: A program evaluation*. Symposium presentation at the annual meeting of the Association of Behavioral and Cognitive Therapies, Toronto, Canada.

Arnold, C. C. & Southam-Gerow, M. A. (2011, November). *The Strength of Measure (SoM) Scale: The development of a systematic method for evaluating the strength of psychological assessment measures within clinical practice*. Poster presentation at the annual meeting of the Association of Behavioral and Cognitive Therapies, Toronto, Canada.

Arnold, C. C. Rodriguez, A., Southam-Gerow, M. A. & McLeod, B. D. (2011, November). *Initial Reliability of the CBT for Youth Anxiety Therapist Adherence Scale*. Poster presentation at the annual meeting of the Association of Behavioral and Cognitive Therapies, Toronto, Canada.

Chin, C. E., **Arnold, C. C.**, & Carter, C. M. (2010, June). *Using Family Literacy Environment and Behavior to Predict Early Language and Literacy*. Poster presentation at the annual meeting of the Head Start Research Conference, Washington, D.C.

Arnold, C. C., Comtois, K., McDonnell, M., Hendricks, K. E., & Morgan, A. (2009, November). *Differences in Psychopathology Among Adolescents Exposed to Violent Trauma, Non-violent Trauma, and No Trauma*. Poster presented at the annual meeting of the Association of Behavioral and Cognitive Therapies, New York, NY.

Arnold, C. C., Quinoy-Boe, A., Brown, R. C., Chu, B. C., McLeod, B. D., Weisz, J. R., Southam-Gerow, M. A. (2009, November). *Therapist Attitude Toward and Use of Manualized Treatments*. Poster presented at the annual meeting of the Association of Behavioral and Cognitive Therapies, New York, NY.

Hourigan, S. E., Quinoy-Boe, A., Brown, R. C., **Arnold, C. C.**, & Southam-Gerow, M. A. (2008, November). *Examining Differences Between Clinic-Referred Youth and Non-Referred Youths on the Children's Emotion Management Scales.* Poster presented at the annual meeting of the Association of Behavioral and Cognitive Therapies, Orlando, FL.

Storck, M. & **Arnold, C. C.** (2007, March). *State of the Art of Adventure-Based Therapies at CSTC and Beyond.* Hosted by Child Study and Treatment Center, Lakewood, WA. Continuing medical education credit awarded through Seattle Children's Hospital, Seattle, WA

TEACHING EXPERIENCE

8/08-5/09	Graduate Teaching Assistant PSYC-317: Research Methods in Psychology Supervisors: Vicky Shivy, Ph.D., Robert Hamm, Ph.D.

CONTINUING EDUCATION

October 29-30, 2007	*Coping With Chaos: Treating Severe Disorders with Dialectical Behavior Therapy* Behavioral Tech, LLC, Portland, OR Marsha Linehan, Ph.D. & Amy Wagner, Ph.D.

PROFESSIONAL MEMBERSHIP

2008-Present	Association for Behavioral and Cognitive Therapies (ABCT)
2010-Present	American Psychological Association, Division 53: Clinical Child Psychology

---

HONORS, AWARDS, SCHOLARSHIPS, GRANTS, AND ACADEMIC SOCIETIES

2011 **Graduate Student Travel Grant**, VCU Graduate School
($120 to present at the Association for Behavioral and Cognitive Therapies 2011).

2011 **VCU Psychology Department Travel Award**, VCU Dept. of Psychology ($350 to present at the Association for Behavioral and Cognitive Therapies 2011).

2006 **College of Liberal Arts and Sciences Outstanding Undergraduate Award**

2006 **Nell Fahrion Outstanding Senior in Psychology Award**

2005 **Golden Key International Honor Society**

2005 **Samuel Priest Rose Memorial Scholarship** for outstanding student in psychology

2004-2006 **Dean's List Scholarship Award**

2004-2005 **CU Denver Academic Achievement Award**, recognition for 4.0 GPA in consecutive semesters

2004 **Psi Chi, National Honor Society in Psychology**