CATEGORICAL AND DIMENSIONAL PRESENTATIONS OF INFORMATION IN CLINICAL FEEDBACK: THE ROLE OF COGNITIVE FIT

By

Mary Michele Athay Tomlinson

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

May, 2013

Nashville, Tennessee

Approved:

Leonard Bickman, Ph.D.

James H. Steiger, Ph.D.

Sun-Joo Cho, Ph.D.

Thomas M. Smith, Ph.D.

To God, for without Him, nothing is possible

and

To my husband, Brooks, for always making me laugh

ACKNOWLDEGEMENTS

This dissertation is the culmination of my graduate school journey and I am grateful for all who have helped, encouraged, and taught me along the way. In particular, I am profoundly indebted to my advisor, Dr. Leonard Bickman, for providing me with incredible research experiences and for sharing his wisdom and knowledge. I am also grateful for Dr. Sun-Joo Cho, who was very generous with her time and knowledge and assisted me in countless ways. Many thanks also go to Dr. James Steiger for his guidance, encouragement, and willingness to share his incredible mind with students.

None of this work would have been possible without the love and support of my family who lifted me up, cheered me on, and always let me know how much they believed in me. I am especially grateful to my loving husband, Brooks, for taking over household chores so I could concentrate only on writing, and for always knowing how to make me laugh. You can still be in charge of cooking and cleaning after I graduate. And to Lauren, my Twin B, who has been by my side literally since the beginning of my life, I am grateful to get to journey through life with you. I could not do it alone.

I am also thankful for the endless support and love I received from amazing people such as Debbie Rankin, Elizabeth Rice, Cheri Hoffman, Susan Haist and Jennifer Samson and would be lost without the unconditional love I receive from my “munchkins” (Caleb, Samuel, Isaac, Isa) and the open invitation for playtime. Finally, I remain forever grateful to Mercy Ministries, which tirelessly works to release young women from

bondage and helps them see their incredible value and worth. I certainly would not be who I am today had I not walked through Mercy's front doors.

Above all else, I am eternally thankful to God for His unwavering and endless love and the peace I have from knowing that He always follows through on His promises. Jeremiah 29:11.

TABLE OF CONTENTS

Page

LIST OF ABBREVIATIONS

Abbreviations

ADHD	Attention deficit hyperactivity disorder
BIC	Bayesian Information Criterion
CBCL	Child Behavior Checklist
CD	Conduct disorder
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
DIF	Differential Item Functioning
DSM	Diagnostic and Statistical Manual of Mental Disorders
GHM	Generalized Mantel-Haenszel Statistic
GRM	Graded Response Model
IRT	Item Response Theory
LCA	Latent Class Analysis
LMR-LRT	Lo-Mendell-Rubin Likelihood Ration Test
LR	Logistic Regression
LTA	Latent Transition Analysis
OCI	Observed Conditional Invariance
ODD	Oppositional defiant disorder
SFSS	Symptoms and Functioning Severity Scale
UCI	Unobserved Conditional Invariance

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

Overview

Clinical judgment and decision-making are critical to psychological practice, yet little research investigates the underlying process of how clinicians make decisions or how these decisions affect clinical outcomes (Gambrill, 2005; Ridley & Shaw-Ridley, 2009). In order to improve mental health outcomes, researchers have developed tools aimed at enhancing clinical decision-making. One such tool supported by empirical evidence is the use of ongoing feedback (e.g., Bickman, Kelley, Breda, Vides de Andrade, & Riemer, 2011; Harmon et al., 2007).

The theory behind the use of ongoing feedback for improving performance has been developed and extensively researched in several fields (see meta-analysis by Kluger & DeNisi, 1996). At first glance, the basic premise for using feedback in mental health treatment is straightforward; if clinicians or other mental health professionals receive accurate ongoing information about a client's treatment process (i.e., therapeutic alliance, motivation for treatment) and progress (i.e., symptoms and functioning, life satisfaction, counseling impact), they can be more responsive to the needs of the client by continuing, discontinuing, or altering treatment plans. This information is often gathered by the use of standardized measures that is then analyzed and reported to the clinician in a feedback report. The actual mechanism by which feedback influences clinical practice is largely

unknown, but is likely a complex process involving numerous factors that include clinician characteristics (e.g., clinician's ability to interpret feedback), institutional policies (e.g., requirement to incorporate feedback in treatment planning), and feedback qualities (e.g., speed and accuracy of feedback). The current dissertation concentrates on one of these factors: how the format of the information in the feedback affects the cognitive processes involved in clinical decision-making.

There remains an ongoing discussion in clinical psychology concerning whether the latent constructs assessed by clinical outcome measures (e.g., psychopathologies) are conceptualized as discrete (i.e., categorical) disorders or continuously distributed (i.e., dimensional) traits, or a combination of both (e.g., Widiger & Samuel, 2005; Witkiewitz et al., 2013). This discussion has become even more energized with the pending release of the fifth edition of Diagnostic Statistical Manual (DSM-5) which incorporates a dimensional conceptualization into its historically categorical system (e.g., Coghill & Sonuga-Barke, 2012; Jones, 2012; Narrow & Kuhl, 2011). However, the assumed structure of the latent construct affects the presentation of clinical feedback. For example, a categorical approach yields information about differences between distinct groups of individuals, whereas a dimensional approach yields information about individual differences in terms of degree. Therefore, the overarching goal of this dissertation is examine how clinical decision-making may be affected by the presentation of information in feedback created from outcome measures assuming the latent construct is discrete or continuous. This adds to the ongoing discussion concerning the structure of latent clinical constructs by offering a comparison in terms of clinical utility for decision-making.

There are four specific aims of this dissertation. The first aim suggests utilizing cognitive fit as a means to compare the clinical utility of the presentation of information. Deeply rooted in information processing theory, cognitive fit relates to how well the format or presentation of information matches with the task being completed (Vessey, 1991). This concept posits that when the information presentation matches with the type of task, the decision-maker is most effectively and efficiently able to apply the information, thus resulting in better decision-making (Vessey, 1991). Review of the literature revealed no prior work modeling the influence of feedback on clinical decision-making via cognitive fit. Therefore, the first aim is to present cognitive fit as a framework by which to explain and examine how the presentation of feedback affects clinical decision-making.

The second aim is to demonstrate how the assumption about the structure of the latent clinical variable made by the statistical model applied to data affects the model output used for clinical feedback. Specifically, statistical models assuming a categorical and dimensional latent variable structure will be applied to cross-sectional and longitudinal clinical data from a large sample of clinically referred youth aged 11 – 18. Applying these models to the same clinical data will allow for the comparison of output information (i.e., what is presented in feedback) across individuals and groups to highlight how different presentations of clinical information are produced from the same data as a result of how the latent clinical variable structure is modeled. This comparison will allow for the observation of similarities, differences, consistencies, and inconsistencies in the nature of the output across the sample. This type of comparison has yet to be done.

Because application of the models from aim two provide different presentations of the clinical information as outputs (i.e., categorical or dimensional), understanding how the specific presentation of information affects the clinical decision-making process is important. This leads directly to the third aim, which is to bring together the first and second aims by integrating the model outputs resulting from aim two with the theory of cognitive fit proposed in aim one. Based on cognitive fit theory, it is suggested that the most effective and efficient decision-making occurs when the presentation of the information as feedback is matched with the specific clinical decision or judgment being made (Vessey, 1991). This will be illustrated using specific case examples from the statistical application described in aim two and applying results from informal clinician surveys concerning how useful different information presentations (i.e., categorical versus dimensional) are perceived for supporting specific clinical decisions and judgments common to youth psychotherapy. In summary, the third aim is to illustrate, based on the clinical application, how cognitive fit proposes to match certain information presentations with specific clinical tasks categorized as categorical or dimensional by informal clinician surveys.

Finally, the fourth aim is to propose future research that may better elucidate the process by which clinicians use information to make clinical decisions and to understand how the specific presentation of information in feedback affects this process. This research will be proposed based on results from the previous aims and will include discussion of a potential future application to ongoing research regarding the effectiveness of a measurement feedback system. The results of the current work, as well as future research, will provide valuable information for assessing the clinical utility of

modeling latent clinical constructs as categorical or dimensional in structure. It will be proposed that model selection be done when considering the clinical impact the latent variable structure has on clinical decision-making.

In summary, the four aims of this dissertation are:

1. To apply a general model of decision-making to clinical decision-making and discuss the role of *cognitive fit* for determining the most effective presentation of clinical information in feedback.
2. To compare model output resulting from the application of statistical models to cross-sectional and longitudinal clinical data that assume different latent variable structures (i.e., categorical and dimensional).
3. To utilize case examples from aim two, as well as information from informal clinician surveys, to demonstrate how the concept of cognitive fit proposes that certain presentations of clinical information support more effective and efficient decision-making based on the specific clinical judgment or decision being made.
4. To propose future research to further investigate the role of cognitive fit for informing how the assumed latent variable structure in a statistical model influences clinical feedback and decision-making.

In addition to describing what this dissertation aims to do, is also important to note what it will not do. This dissertation will not provide a definitive conclusion concerning how feedback influences the clinical judgment process. Nor will it offer a specific conclusion about how clinical information should be presented in feedback.

Such conclusions are only possible with additional research, some of which will be proposed in the discussion. However, this dissertation provides a foundation for future work and introduces a novel idea for approaching this discussion.

In conclusion, the overall goal of this dissertation is to provide a source of information to consider when weighing in on the debate concerning how to conceptualize the structure of latent clinical variables. This dissertation will suggest it is important to consider how feedback that results from each conceptualization affects clinical decision-making. It will be concluded that perhaps a latent clinical variable should be modeled based on the structure that results in a presentation of clinical information that most enhances clinical decision-making. Ultimately, this may lead to improved client outcomes.

Clinical Decision-Making and Feedback

Prior to introducing the data and methods used in the clinical application portion of this dissertation, a more detailed discussion of available literature will be presented, supporting the aims presented in the previous section.

Clinical Decision-Making

Clinical decision-making is an integral part of the psychological treatment process. Clinicians are continuously making decisions such as whether or not to treat a client, how often to meet with a client, what diagnosis (if any) should be made, what type of treatment to use, what the treatment goals are, the appropriate level of care (i.e.,

outpatient, inpatient, residential, etc.), whether referral for medication or to another provider is needed, etc. Indeed, as Ridley and Shaw-Ridley (2009) stated, "Clinical judgment is foundational because positive therapeutic outcomes hinge on establishing reasonable goals and selecting appropriate treatments, and appropriate treatment selection hinges on sound judgment and accurate decision making" (p. 401). In this way, the effectiveness of the treatment process is highly dependent on the ongoing validity of the clinician's picture of the client.

Clinicians are constantly gathering information (informally and/or formally) and assessing a client's status and context in order to create a valid picture of the client's psychological needs. However, this picture is not static. Instead, it changes over time and, to keep it valid, a clinician must continuously gather information and complete assessments in order to maintain an accurate picture that directs treatment planning. In painting this picture, however, clinicians are faced with the task of integrating and interpreting multiple sources of ongoing information concerning client's idiographic experiences (e.g., symptom patterns, overt behaviors, and covert personality dynamics) within the context of the specific social, cultural, and environmental factors (Ridley, Tracy, Pruitt-Stephens, Wimsatt, & Beard, 2008). The ability to obtain information and then accurately integrate and interpret such information directly influences the validity of the clinical picture and therefore the resulting clinical decisions and treatment outcomes. In this way, "clinical decision-making truly is at the heart of clinical practice" (Gambrill, 2005, p. 7).

Despite the importance of clinical decision-making on clinical outcomes, there remains a surprising dearth of research on the underlying process by which it works

(Falvey, Bray, & Herbert, 2005; Stewart & Chambless, 2008), and subsequent calls for research have been made (Kazdin, 2008; National Institute of Mental Health, 1999; Puschner et al., 2010; Street, Niederehe, & Lebowitz, 2000; Willis & Holmes-Rovner, 2006). The two topics of research currently represented in the literature concerning clinical decision-making include investigating strategies clinicians used for making clinical decisions and the continuing 50-year debate concerning the accuracy of actuarial versus clinical prediction.

Research on strategies clinicians use for making clinical decisions is largely descriptive and based on clinician self-report. For example, Steward and Chambless (2007, 2010), asked clinicians to report the strategies they used to determine treatment goals for a client who was getting worse. Overall, clinicians reported greatest utilization of past clinical experiences when making treatment decisions. Similarly, in rating the importance of various sources of information, the importance placed on past clinical experience was significantly higher than importance placed on using current research, discussions with colleagues, and experiences in personal therapy. Interestingly, there was no difference in reported reliance on current research compared to the clinician's own personal experiences receiving therapy. In another study, Stewart and Chambless (2008) utilized a slightly different methodology and asked clinicians to recall a difficult case they encountered where the client was not progressing. They were then asked how they had proceeded in treating this client, allowing for the use of multiple strategies. Eighty-five percent of clinicians reported consulting with a colleague and 76% reported relying on their past clinical experience for determining the next steps to take. Only 41.6% of

clinicians reported consulting with current research and 10.2% of clinicians did nothing different and continued with the client with what they were doing.

Although only a few examples were provided here, descriptive research suggests a consistent reliance on clinical experience over other techniques or sources of information for clinical decision-making, including referring to current research. This is a long-standing trend in the field. As early as 1986, Morrow-Bradley and Elliott found that approximately half of clinicians surveyed indicated past clinical experience as the most important source of information and only 10% reported using psychotherapy research sources. Garb (2005) articulated this trend when she concluded that, like other human decisions, clinical judgments are often based on personal experiences and other intuitive processes rather than empirical reasoning.

The other vein of research investigating clinical decision-making dates back over half a century and assesses the accuracy of actuarial (i.e., statistical) versus clinical prediction. Such research compares whether clinicians or mechanical procedures (i.e., statistical or actuarial techniques) better predict human behavior, including diagnoses, prognoses or assess states and traits (e.g., abnormal behavior and personality; Grove, Zald, Lebow, Snitz & Nelson, 2000). For example, researchers have tested how accurate clinicians are at identifying clients at risk for treatment failure (i.e., drop out or clinical deterioration) compared to the accuracy of statistical techniques for identifying them. This literature has confirmed, numerous times, the general superiority of statistical prediction over clinical judgment (e.g., Dawes, Faust, & Meehl, 1989; Grove et al., 2000; Lutz, Lambert, Harmon, Tachitsaz, Schürch, & Stulz, 2006; Meehl, 1954; Sawyer, 1966). To summarize this literature, Grove et al. (2000) conducted a meta-analysis of studies

that compared clinical versus mechanical prediction and found that, on average, mechanical predictions were about 10% more accurate. In fact, in nearly half of all studies included, mechanical prediction was substantially more accurate compared to clinical prediction.

Although there has been a long-time push by researchers for the inclusion of statistical prediction techniques into clinical practice, it is often met with resistance from practitioners (Bell & Mellor, 2009). Therefore, despite research consistently finding the benefits of statistical prediction, the debate is far from over. In fact, this debate will only intensify with the advancement in computer technology and software programming that allows for the organization of large amounts of data to make complex decisions (e.g., Garb, 2000). However, as expressed by Spengler, White, Ægisdóttir, and Maugherman (2009), "although it may be difficult to convince counseling (and other) psychologists to utilize actuarial techniques, findings from client outcome and other emerging research programs suggest actuarial feedback may improve clinical decision-making accuracy" (p.419). This reflects a shift the debate has taken from arguing for one method over the other to focusing on how clinical and statistical techniques can be used together to enhance clinical practice (Bell & Mellor, 2009; Falzer & Garman, 2012; Kazdin, 2008; Snyder, 2000). Perhaps one of the best examples of integrating statistical techniques with clinical decision-making is the use of systematic evaluation (i.e., outcome assessment) to create feedback for use in treatment planning and monitoring. This will be further discussed in the following section.

Clinical Feedback and Decision-Making

Clinicians are central to the therapeutic process. Their clinical judgments and decisions shape the course of treatment, which has great implications for client outcomes. Therefore, improving clinical care depends on enhancing the clinician's level of accuracy in making judgments and decisions (Bell & Mellor, 2009; Holt, 1986; Snyder, 2000). One proposed and well-studied way of doing this is through the use of ongoing clinical feedback. In fact, the American Psychological Association (APA) Presidential Task Force on Evidence-Based Practice (EBP) recommended greater use of outcome monitoring and feedback for all practices (APA Presidential Task Force on Evidence-Based Practice, 2006). Worthen and Lambert (2007) suggest that such feedback influences clinical outcomes by providing information clinicians may have unintentionally overlooked or underemphasized and by identifying problems within specific domains which may jeopardize treatment completion or progress. And, in fact, research demonstrates the use of clinician feedback improves clinical outcomes in mental health treatment (Anker, Duncan, & Sparks, 2009; Bickman et al., 2011; Harmon et al., 2007; Lambert, Harmon, Slade, Whipple, & Hawkins, 2005; Reese, Norsworthy, & Rowlands, 2009; Slade, Lambert, Harmon, Smart, & Bailey, 2008).

Although feedback may include a variety of indicators such as therapeutic alliance, treatment motivation, and life satisfaction (e.g., Bickman et al., 2011), the most common form of clinician feedback in psychotherapy research includes indicators of patient progress regarding psychopathology, which are meant to alert the clinician to whether the client is improving, deteriorating, or not changing. Currently, the client's symptom severity (one of the primary constructs used as a clinical outcome) is most

frequently measured for the purpose of clinical feedback (Shulte, 1997). The measurement of symptom severity then provides feedback the clinician can use to revise treatment planning, alter the focus within a session, change treatment modalities, or increase/decrease the level of care. For example, Lambert and colleagues (2005) utilized the Outcome Questionnaire-45 (OQ-45; Lambert, Gregerson, & Burlingame, 2004) to assess clients' clinical symptoms and functioning. In creating feedback, Lambert et al. used a Reliable Change Index (RCI) as well as clinical and normative data to indicate if clients: 1) changed in a clinically meaningful way; and 2) displayed scores typical of a dysfunctional or functional population. Feedback based on these two indicators was given to clinicians regularly with the purpose of guiding their treatment planning (e.g. termination or altering of treatment plan). The provision of this feedback has been shown to be effective in improving clinical outcomes for adults receiving mental health care in university settings (Lambert et al., 2002, 2005). Feedback was particularly effective for clients who were deteriorating during the course of treatment.

Bickman and colleagues (2011) have also demonstrated the effectiveness of providing clinicians ongoing feedback for improving client outcomes. In their study, the Symptoms and Functioning Severity Scale (SFSS; Bickman et al, 2007, 2010) was administered regularly to clients (youth aged 11 – 18) receiving mental health treatment. Parallel forms of the SFSS were also administered to the clients' caregivers and clinicians. Information gathered from these measures was entered into a computerized system (Contextualized Feedback Systems®, CFS™) that generated immediate feedback reports for clinicians. This feedback included indication of whether the youth had: 1) high, medium, or low symptom severity compared to clinical norms; and 2) displayed

meaningful change (improvement or deterioration) from previous measurement based on an index of minimum detectable change (MDC; Schmitt & Di Fabio, 2004). In a cluster randomized experimental design to detect the effect of feedback on youth outcomes, youth whose clinician received ongoing feedback improved faster than youths treated by clinicians not receiving feedback (Bickman et al., 2011).

Despite the increasing evidence supporting the use of feedback to improve clinical outcomes, little is known about how clinicians use this information to make decisions and how effective these decisions are (Falvey et al., 2005). However, it is generally assumed that the provision of feedback enhances clinical decision-making. Yet, given the unobservable nature of psychological phenomena, researchers must rely on manifest (observable) variables in order to infer about the underlying latent (unobservable) constructs they wish to provide feedback on. As seen in the studies conducted by Lambert et al. (2002, 2005) and Bickman et al. (2011), the feedback provided to clinicians was based on information gained from the repeated administration of standardized outcome measures. In this way, the feedback is intimately tied to any assumptions made about the latent construct (i.e., symptom severity) being assessed by the outcome measure. One assumption that is made when analyzing such data concerns the underlying structure or form of the latent clinical construct (i.e., whether it is categorical or dimensional in nature). This assumption affects the nature of the feedback used by clinicians and therefore potentially their clinical decision-making. The following section provides a discussion about the assumptions underlying the form of latent constructs, specifically when they are categorical or dimensional in nature.

Assumptions about the Form of Latent Constructs

Clinical psychology is concerned with psychopathologies, which, while often described by observable behaviors (i.e., symptoms), are actually unobserved (latent) constructs. These latent clinical constructs can take on several different structural forms, two of which are categorical and dimensional. This section will describe and illustrate examples of these two structural representations of latent variables applied to the assessment of psychopathology.

Latent Variables as Categorical Constructs

Latent variables as categorical constructs are composed of discrete and mutually exclusive groups displaying within-group homogeneity. For example, individuals can be either "clinically depressed" or "not clinically depressed," depending on manifest responses to specific depression symptom indicators/items. These classes are distinctly different from each other and within each respective group, each individual is assumed to have the same predicted probability of endorsing each specific depression symptom. For example, all individuals in the 'clinically depressed' class would have the same predicted probability of endorsing the item '*difficulty getting out of bed in the morning*.' Additionally, within a class, each specific depression symptom holds equal 'weight' for classification in the class. For example, endorsing the item '*thoughts of suicide*' carries no more or less weight for classifying an individual in the clinically depressed class than does endorsing the item '*loss of interest in activities*.'

Perhaps the best example of a categorical approach in psychology, as well as the most dominant view of psychopathology, is demonstrated with the *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV;* American Psychiatric Association, 2000). The DSM-IV lists explicit criteria needed for an individual to be placed in one of a large number of specific and categorical mental disorders (Krueger, Watson & Barlow, 2005). These categorical diagnoses are nearly always required for reimbursement by insurance companies (Aboraya, 2007; Chodoff, 2002; Hatfield & Ogles, 2004) and have been called an "admission ticket to eligibility for mental health services and reimbursement" (Bickman, Wighton, Lambert, Karver, & Steding, 2012, p. 1). With the DSM, individuals either meet the relevant criteria for a specific diagnosis or they do not – few fall in the "grey" area between these categories (Grayson, 1987). This approach assumes that at that diagnostic cut-off (or threshold of a symptom count) where an individual exceeding the cut-off is diagnosed with a specific disorder, there is a sharp increasing probability of having that disorder. In other words, an individual displaying one less than the required threshold number of symptoms has a low probability of having the disorder but an individual displaying the threshold number of symptoms has a high probability of having that particular disorder. These individuals would be considered distinctly different from each other even though there is technically only one symptom differentiating them.

Latent Variables as Dimensional Constructs

Alternatively, psychological constructs can be viewed as continuous in nature. These constructs are often called *dimensional* although *continuous* is more accurate given

that categorical approaches also technically have dimensions (they are discrete). This distinction is not generally recognized in the literature and *dimensional* is used interchangeably with *continuous*. This paper will follow the tradition of the literature and use the term dimensional. Dimensional (or continuous) constructs are measured on a latent continuum. Individuals are placed at different points on the continuum depending on their scores on manifest attributes. Thus, instead of two homogeneous classes (e.g., clinically depressed vs. not), the dimensional view places all individuals somewhere along the continuum from "not at all clinically depressed" to "severely clinically depressed". As a result, individuals have different predicted probabilities of endorsing each symptom, depending on their level of the latent trait (i.e., depression) and the level of severity each item represents on the continuum. For example, individuals who are severely clinically depressed would have a higher predicted probability of reporting suicidal ideation (an item that indicates high clinical severity) than those who are less clinically depressed. Similarly, endorsement of the item '*feeling complete hopelessness and despair*' might indicate a higher level of clinical depression than endorsement of the item '*feeling a little sad*' and thus a person with more severe clinical depression would have a high predicted probability of endorsing the former.

Dimensional approaches are often used when assessing the degree or magnitude of illness. Therefore, it is common for clinicians and researchers to think in terms of a dimensional perspective when assessing clients' change over time or are in need of diagnosis-specific quantitative scores (Helzer, Kraemer, & Kruegar, 2006). In this way, the number or pattern of symptoms is viewed on a continuum. The dimensional approach assumes that clinically significant individual differences exist among individuals who fall

above or below a categorical diagnostic threshold (Helzer et al., 2006). In this case, psychopathology would be described by degree. For example, the individual with one fewer symptom than a threshold number would not be distinctly different from an individual with the threshold number, but rather simply display some reduced degree of that disorder. Examples of dimensional approaches currently used in clinical psychology and psychiatry can be seen with the use of measures such as the Hamilton Scale for Depression (Hamilton, 1960), the Child Behavior Checklist (CBCL; Achenbach, 1985) and the corresponding Youth Self-Report (YSR; Achenbach, 1991), and the Positive and Negative Syndrome Scale for schizophrenia (PANSS; Kay, Fisbein & Opler, 1987), when individual scores are computed as the sum or average of item responses.

The next section will present some issues specific to clinical measurement when latent clinical variables are conceptualized as categorical or dimensional. One goal of the following section is to illustrate some of the issues brought up between researchers and practitioners in the long-standing debate concerning whether latent psychopathologies are described as discrete (i.e., categorical) or continuous (i.e., dimensional) traits.

Specific Issues in Clinical Measurement

The debate between categorical versus dimensional conceptualizations of latent clinical variables has been around a long time and will not be resolved in the near future. There is increasing evidence both in favor of and against each perspective within clinical psychology. It is possible that the *true* structure of these unobservable psychopathologies in the universe will never been known. However, when analyzing data from clinical

outcome measures, it is important to recognize that the statistical model used makes an assumption about this structure. This section will highlight some of the general advantages, disadvantages, and difficulties seen in measurement and application when latent clinical constructs are analyzed as either categorical or dimensional.

Categorical Approach

Much of the scientific research in clinical psychology is based on the categorical system designated in the DSM-IV. This approach has several advantages. For example, the development and widespread use of the DSM-IV system for categorizing psychopathology has been beneficial for classifying and labeling mental disorders and in facilitating communication among researchers and clinicians (e.g., Mack, Forman, Breown & Frances, 1994). Additionally, the categorical system provided by the DSM-IV has been said to provide clear direction for clinical decision-making (e.g., to prescribe medication or not; to hospitalize or not) and is often the basis of reimbursement policies used by insurance companies (Kamphuis & Noordhof, 2009). Despite these recognized advantages for applying a categorical approach, there are many drawbacks, several of which raise questions about whether or not this approach is sufficient and valid for representing latent clinical constructs of psychopathology. Two main features assumed within a categorical approach often fall short when applied to the widely used categorical DSM-IV. These are: (1) homogeneity within category, and (2) definitive boundaries between categories resulting in qualitatively different groups.

First, given the assumptions of the categorical perspective, one would expect to find great similarity (i.e., homogeneity) between all individuals within a specific group.

It is assumed that the individuals who share a specific diagnostic category also share at least some common features, symptoms, and levels of psychological functioning. Furthermore, they would be quite dissimilar to those who are not within the diagnostic category. Unfortunately, decades of clinical application have demonstrated the DSM has difficulty meeting these expectations. For example, within various diagnoses of personality disorders (e.g., schizotypal, borderline, narcissistic), the diagnostic criterion includes meeting five out of nine listed symptoms for formal diagnosis. Thus, theoretically, it is possible for two individuals with the same diagnosis to have only one symptom in common. Similarly, as Krueger, Watson et al., (2005) pointed out in the case of obsessive-compulsive disorder, it is theoretically possible for two diagnosed individuals to have no features in common. These examples demonstrate there can be great heterogeneity within any given diagnostic category, challenging an assumption underlying the categorical perspective.

The second assumption of the categorical perspective that often fails to be upheld in practice is that of distinct boundaries between categories. This assumption implies that each category is qualitatively different from the others and that individuals clearly belong to a category or they do not. Unfortunately, comorbidity and subthreshold disorders have all been cited as evidence that psychopathology has difficulty meeting this categorical assumption (Krueger, Watson et al., 2005; Maser et al. 2009; Widiger & Samuel, 2005). First, comorbidity, or the coexisting of distinct DSM-IV disorders, proves to be difficult to explain from a categorical perspective (Maser et al., 2009; Widiger & Samuel, 2005). Research suggests that over 50% of individuals with one psychiatric diagnosis also have at least one other psychiatric diagnosis (Kessler et al., 2005). For example, in their 2008

study of comorbidity patterns in children and adolescents, Elia, Ambrosini, and Berrettini found that over 40% of youth diagnosed with attention deficit hyperactivity disorder (ADHD) were also diagnosed with oppositional defiant disorder (ODD). Similarly, Ford, Goodman, and Meltzer (2003) found in an epidemiological study that over half of children diagnosed with ADHD had a comorbid behavior disorder (ODD, conduct disorder, or another disruptive disorder). Thus, comorbidity appears to be the rule and not the exception. In fact, for some diagnoses, one of the best predictors for meeting the criteria for the diagnosis is whether criteria are met for another diagnosis (Kreuger, Markon, Patrick, & Iacono, 2005). This systematic co-occurrence of diagnoses calls into question whether or not these are distinctly different categories or whether they are really variations of the same disorder/dysfunction. As another example, although anorexia nervosa and bulimia nervosa are separate diagnostic categories in the DSM-IV, their high level of comorbidity and frequent diagnostic crossover in patients may indicate that they are not so distinct (Eddy, Dorer, Franko, Tahilani, Thompson-Brenner & Herzog, 2008).

The second reason why the assumption of distinct boundaries is questioned within the categorical system of the DSM-IV is the presence of sub-threshold psychopathology. If a specific psychopathology were a purely categorical construct, individuals would fall either 'in' a category or 'out' of it. In other words, there would be a distinct separation between categories, as well as contrasting between-group predicted probabilities of having specific symptoms. Furthermore, few individuals would be in the 'grey' area between categories. Unfortunately, much research has found that individuals who are sub-threshold (i.e., have symptoms just below the clinical threshold) have significant distressing and clinically meaningful psychopathology that may be obscured by the

reliance on distinct category boundaries (e.g., Judd, Akiskal & Paulus, 1997; Judd, et al. 2008; Okasha, 2009). For example, in a study of how subthreshold alcohol dependence predicted future outcomes, individuals not meeting the diagnostic threshold for alcohol dependence but who still manifested subthreshold levels, were at a significantly increased risk for developing more severe alcohol problems later on (McBride & Adamson, 2010). Additionally, Okasha (2009) reported no significant differences between groups of individuals with a full posttraumatic stress disorder (PTSD) diagnosis and those with subthreshold PTSD in regards to degree of impairment. Based on evidence such as this, a categorical system may put those with subthreshold disorders at a severe disadvantage (Goldberg, 2000; Shankman, Lewisohn, Klien, Small, Seeley & Altman, 2009). In fact, many argue that this calls into question whether psychological disorders are truly categorical as the DSM-IV portrays them to be.

Dimensional Approach

The dimensional conceptualization of psychopathology also has advantages and disadvantages. An advantage to a dimensional perspective is the ability for ongoing tracking of psychopathology and monitoring of symptoms throughout treatment. In this way, even small changes can be interpreted as meaningful. In fact, it appears that clinicians naturally think with a dimensional structure when characterizing their patients' psychiatric status and progress (Maser et al., 2009). Assessing a client's DSM-IV diagnostic status continually (i.e., through a categorical perspective) provides little information for clinicians with regard to the client's progress or whether a particular treatment is benefiting them, unless a client moves from a clinical (diagnosis-present) to a

non-clinical (diagnosis-absent) status (Helzer et al., 2006). A second advantage to the dimensional perspective is that it can provide a quantitative score using a consistent methodology that can easily be compared across studies or individuals. This enables clinicians to communicate, at a quantitative level, about clinically relevant information such as symptom severity or physical impairment (Helzer et al., 2006). Finally, the dimensional approach may better reflect the prevalence of comorbidity and subthreshold disorders previously described as drawbacks to the categorical conceptualization of psychopathology. Dimensional approaches may help ensure that clients are treated for their full range of psychopathology, therefore producing better outcomes (Helzer & Hudziak, 2002).

Although the dimensional system boasts many advantages, there are some challenges inherent in the dimensional conceptualization. For example, a clinician can easily communicate via diagnostic labels but for some clinicians, communication may become more complex with quantitative or continuous variables must be communicated, even if the information is more specific and precise (Helzer al., 2006; Widiger & Samuel, 2005). Furthermore, some think that diagnostic categories also lend themselves more easily to some clinical decisions (e.g., to hospitalize or not; to treat with medication or not). Therefore, a limitation of a dimensional structure is that clinicians may need additional guidance for making decisions with dimensional information. Unfortunately, no consistently applied or well-understood threshold for the presence or absence of a categorical diagnosis currently exists within a dimensional framework (Widger & Samuel, 2005). Therefore, the DSM-IV is still the 'gold-standard' for assigning a diagnostic label. A related disadvantage to a dimensional perspective is that diagnostic

labels are used for granting insurance coverage and billing purposes. Given the lack of universally applied thresholds in a dimensional approach for diagnosing, it is unclear how a dimensional system might be utilized for these purposes.

An additional challenge in the dimensional approach for measurement of psychopathology is that many clinical scales used to measure psychopathology are often developed either for clinical (pathological) use or for "normal" populations. Therefore, the items on a given measure often all represent either a normal range or a high range of psychopathology, providing information only within that respective narrow range of the dimension. Thus, a full picture of the spectrum is not always represented. Understanding the entire spectrum from severe levels of psychopathology to the absence of psychopathology is needed to provide complete information about the construct, as well as indicate the critical boundary zone between these two extremes that have different clinical needs (Cuthbert, 2005). The use of measures assessing narrow ranges of a dimension also generally result in rather homogeneous item content selected to represent only that range (Reise & Waller, 2009). When this happens, one would only expect items to have a high ability to differentiate between individuals within that narrow range and those not in that range. This gives the measure the appearance of being somewhat more categorical than dimensional in nature, limiting the amount of information provided. Reise and Waller (2009) found, in the majority of clinical measures they reviewed, that items on clinical measures targeted only one portion of the spectrum. In order to capitalize on the advantages provided by a dimensional system, clinical measures with items able to detect more precisely the level of psychopathology along the entire

continuum are needed. However, this may result in lengthy measures that create significant time burden for their completion.

In conclusion, there is no consensus among researchers or practitioners concerning whether psychopathologies are categorical or dimensional latent constructs. Furthermore, representing psychopathologies as either categorical or dimensional latent constructs has potential advantages and disadvantages. Therefore, clinical data from outcome measures can be analyzed under either assumption, depending on the statistical model used. The following section will discuss research findings concerning the conceptualization of externalizing problems as categorical or dimensional in structure, as well as research that has directly compared these different structures. The discussion is limited to externalizing problems in clinically referred youth as it is the focus of the empirical aspect of this dissertation.

Dimensional Versus Categorical Latent Structure of Externalizing Problems

The debate about the latent structure of clinical variables concerns all populations, including children, adolescents, and adults. Because the empirical application of this dissertation concerns youth (here, ages 11 – 18), the current discussion now narrows to discussion of only child and adolescent populations (i.e., under 18 years of age).

Childhood psychopathology and behavior disturbances are often categorized by two major categories: externalizing and internalizing problems (De Clercq, De Fruyt, Van Leeuwen & Mervielde, 2006; Kooijmans, Scheres & Oosterlaan, 2000; Leve, Kim & Pears, 2005). These common descriptors of behaviors that characterize youth mental

disorders have dominated research for numerous decades (Achenbach & Edelbrock, 1978; Van der Akker, Dekovic, Asscher, Shiner, Prinzie, 2012; Woods, Farineau, McWey, 2013) and continue to be reliable clusters of problem behaviors. Thus, measures of symptom severity or functioning often assess behaviors based on these two categories (e.g., CBCL, Achenbach, 1985). Internalizing problems are problems characterized as happening "within the self," such as fears, sadness, physical complaints, worrying, shyness, etc. Internalizing disorders include things such as anxiety and depression. Historically, internalizing disorders have been classified as "neuroses" but have also been called "over-controlled" or "over-inhibited" (Achenbach, 1985). Children seem to deal with their problems internally, rather than acting them out in the environment. On the other hand, externalizing problems are characterized by disruptive behaviors that are directed outward, typically toward other people and involve conflict. Examples include disobedience, aggression, delinquency, temper tantrums, and over-activity. Externalizing disorders include ADHD, ODD, and conduct disorder (CD). Externalizing disorders are often called "under-controlled" or simply "aggressive" because children act out their problems externally (Achenbach, 1985).

Although there exists wide acceptance for the definitions of and the division between internalizing and externalizing problems, there is not yet a consensus on the underlying structure of these constructs. Specifically focusing on externalizing problems, the following sections will discuss research findings when this latent clinical construct was represented as categorical or dimensional structure.

Externalizing Problems as a Categorical Latent Construct

The taxonomy of externalizing behavior in children and adolescents generally contains three problem areas or domains: attention/hyperactivity problems, aggressive/oppositional problems, and delinquent/conduct problems (American Psychiatric Association, 2000). These represent distinct diagnostic categories within the DSM-IV. Therefore, many researchers view externalizing problems as a categorical latent construct and use methods such as latent class analysis (LCA) to investigate externalizing problems. Rather than determining diagnoses or clinical scores in these problem domains by predetermined scores, LCA allows for the identification of relatively homogeneous classes of problem behavior that are different from one another based on item responses or endorsement of symptoms/behaviors. Thus, the goal of LCA is to identify classes of youth according to how individual symptoms patterns naturally occur based on empirical decision rules. Then the relative probability of individuals being assigned to a class can be computed according to a defined set of behavioral referents.

Many studies have used LCA to investigate externalizing behavior in youth. For example, Storr, Accornero, and Crum (2007) utilized LCA with the items on the Youth Self Report related to ADHD, ODD, and CD problem behaviors. They found that three latent classes fit the data best, where the classes varied mainly on severity (i.e., increase in item endorsement probabilities from class to class) but also in frequency of different features. The largest class was a normative class with no clinical features consistent with disruptive behaviors, a second class was considered a class with clinical features of ADHD and ODD (but low CD), and the third class had high clinical features of ADHD, ODD, and CD. Research conducted by deNijs, vanLier, Verhulst, and Ferdinand (2007)

used LCA to investigate patterns of externalizing behavior in adolescents by using items included on the attention problem, aggressive behavior, and rule-breaking scales of the CBCL (Achenbach, 1985). They found six distinct classes of youth. Three of the classes displayed problems in all behavior areas but with differing degrees. A fourth class had mostly attention problems without aggressive behaviors, class five had mild attention problems with low aggressive behaviors, and the last class was the normative group with no externalizing behaviors. Again, this research demonstrated that individuals could be classified into distinct groups based on patterns of behavior across all externalizing problem areas, although they were not always consistent with the DSM classification system.

Evidence for distinct groups has also been found when using LCA to investigate youth with specific diagnoses. For example, Ostrander, Herman, Sikorski, Mascendaro, and Lambert (2008) investigated patterns of psychopathology in children with ADHD and found six distinct classes. Although half of the children fell into classes that could not be reliably distinguished using DSM-IV subtypes, the other groups were distinguishable based on co-occurring symptoms. Lacourse, Baillargeon, Dupere, Vitaro, Romano, and Temblay (2010) also found four distinct classes of youth diagnoses with CD based on the nature, frequency, type of behavior, and co-occurring symptoms. As a whole, these studies suggest that youth externalizing problems can be described by distinct categories with clear boundaries between groups.

Externalizing Problems as a Dimensional Latent Construct

Despite the frequent use of categorical models to analyze externalizing problems, many researchers believe there is little justification for the existence of discrete groups. Instead, they propose that a dimension more accurately depicts externalizing problems, where individuals differ by degree rather than type (Bauer, 2007; Maughen, 2005). In view of this, researchers often use models under the framework of item response theory (IRT) in order to model externalizing symptoms (or items) in terms of their relationship to an underlying latent continuum.

Prior to summarizing results from previous research using IRT to model externalizing problems, it is important to differentiate an IRT approach from an unweighted total score approach in Classical Test Theory (CTT; Novick, 1966; Spearman, 1904). In the CTT total score approach described here, item responses are simply added or averaged to create a total score. Thus, total scores can range from the lowest possible scale score (i.e., when all items are endorsed at the lowest response category) to the highest possible scale score (i.e., when all items are endorsed at the highest response category). The IRT approach, on the other hand, models the response of a person to each item on the measure given their level of severity (i.e., person's latent trait estimate). In other words, within IRT, endorsement of items indicates varying levels of externalizing severity (i.e., items have different locations or "severities"). When a person's trait level is matched with the severity of an item, the person has a predicted probability of .5 for endorsing the item. The linking of each item to a specific level of the latent trait is an advantage of IRT, as it allows for more precise measurement of youth externalizing severity along the continuum. For this reason, IRT is often regarded as

superior to CTT methods. For example, in their direct comparison of CTT and IRT methods for analyzing measures assessing psychopathology with Likert-type items (i.e., where possible responses range from *strongly disagree* to *strongly agree*), Dumenci and Achenbach (2008) recommend using IRT approaches that take into account the ordinal item distributions. This recommendation is made particularly because CTT-sum approaches are biased at the ends of score distributions (see also Wright, 1999).

Prior research using IRT has demonstrated that items often used to assess externalizing disorders or behaviors had different severity levels. For example, in their IRT analysis of items used in diagnostic interviewing for CD based on the DSM-IV, Gelhorn and colleagues (2009) found that the item concerning bullying indicated a higher item severity (i.e., higher item location) than the item concerning lying. Similarly, Kim, Kim, and Kamphaus (2010) found teacher-rated items used on aggression scales of the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) had differing levels of severity. For example, the items '*threatens to hurt others*' and '*bullies others*' had higher item severities than '*argues when denied own way*'. Lambert, Essau, Schmitt, and Samms-Vaughan (2007) conducted IRT analyses on the YSR of the CBCL (Achenbach, 1991) in German and Jamaican adolescents and found similar results. They found items asking about threatening to hurt others or getting into fights had higher item severities compared to items about difficulty paying attention, lying, or disobeying parents. Rapport, LaFond and Sivo (2009) investigated item severities for the aggression and delinquency items of the CBCL in a sample of American boys aged 6 – 16 and also found items asking about threatening people, physically attacking others, and bullying had higher item severities compared with items concerning lying, restlessness,

disobedience, and impulsivity. These studies share the similar result that items referring to aggressive acts towards others (i.e., bullying or threatening others) had higher levels of externalizing severity than less physically aggressive items such as lying or arguing. As a whole, these studies suggest greater precision in assessing a youth's externalizing symptom severity by using IRT, given that items demonstrate differing levels of severity. Therefore, individuals would have different predicted probabilities of endorsing these symptoms based on their clinical severity.

Externalizing Problems: Categorical vs. Dimensional

As with a majority of research that used LCA or IRT to investigate youth externalizing symptom severity, most researchers adopting one particular model over another do not mention consideration of alternative families of models with different latent variable structures. Instead, decisions for assuming a categorical or dimensional structure for externalizing problems represent an a priori preference for one over the other (Klein & Riso, 1993; Widiger & Samuel, 2005). However, several studies have statistically compared models that assume a categorical, dimensional, and even a hybrid (i.e., categorical and dimensional) variable structure. For example, Krueger, Markon, et al., (2005), directly compared categorical versus dimensional conceptualizations of externalizing problems in adults by applying both latent class and IRT models to empirical data. Based on comparison of model fit indices, authors concluded that externalizing problems were 'best' described by a dimensional latent structure. Markon and Krueger (2005) made a similar conclusion in their direct comparison of the relative fit of latent class and IRT models to externalizing behaviors in adulthood.

Currently, only one published study has directly compared the fit of categorical and dimensional models to externalizing problems in youth. In this study, Walton, Ormel, and Kruegar (2011) compared the fit of latent trait (dimensional), latent class (categorical), and factor mixture (hybrid) models to aggression and delinquency indicators from the CBCL in a population-based sample of adolescents. Consistent with results from adult studies, a dimensional structure provided the 'best' fit to the data according to model fit statistics. Thus, in these direct statistical model comparisons, it appears that consensus favors a dimensional structure for modeling externalizing problems for both youth and adult populations. Although psychology and psychiatry have been historically dominated by a categorical system (i.e., the DSM) for describing psychopathology, research findings such as these have increased the push by researcher and practitioners to incorporate a dimensional perspective to the DSM. In fact, the recent revision of the DSM-V includes a dimensional emphasis within its categorical framework (Brown & Barlow, 2005; Coghil & Sonuga-Barke, 2012; Jones, 2012; Kamphuis & Noordhof, 2009; Narrow & Kuhl, 2011)

Despite theoretical debates and statistical comparisons advocating either a dimensional or categorical approach, evidence seems to point to a great overlap between these perspectives. Not only have researchers found that both dimensional and categorical approaches offer useful distinctions for conceptualizing externalizing severity, but the results from analyzing data with these assumptions may be quite similar as a whole. It is interesting to notice that a majority of studies using latent class analysis found that the classes that emerged that were ordered in terms of severity. In other words, the class-specific probabilities of item endorsement increased monotonically

across classes instead of being unique in each class. Thus, the latent classes appear to represent a degree of severity on an underlying continuum rather than distinct groups in a nominal sense. This distinction can be seen pictorially in Figure 1. In panel A, the latent classes are ordered along a continuum where classes are ranked from no severity to high severity. In this case, the probability of item endorsement increases across the classes with individuals in the high severity classes having a higher probability of item endorsement than to those in the mild severity class, who have a higher probability of item endorsement compared to those in the no severity class. Alternatively, panel B depicts latent classes that are distinctly different in a nominal sense. Here, the classes differ based on ODD and CD diagnosis and the probability of item endorsement is strictly unique within a class (i.e., each latent class has a unique set of symptoms or behavioral patterns). In addition the research presented previously in this chapter, numerous researchers have found evidence supporting panel A for representing externalizing symptomology with ordered latent classes (Eaves et al., 1993; Lacourse et al., 2010; Nock, Kazdin, Hiripi, & Kessler, 2006; Odgers, Moretti, Burnette, Chauhan, Waite, & Reppucci, 2007; Storr et al., 2007).

The presence of ordered latent classes in terms of degree of severity is consistent with the dimensional perspective that externalizing severity is described on a continuum. The difference is that the categorical perspective identifies distinct groups in terms of degree of severity (i.e., qualitative differences), whereas the dimensional perspective identifies individual differences in degree of severity (i.e., quantitative differences). This makes it difficult to arrive at a definitive conclusion concerning the true nature of externalizing severity. However, a choice has to be made when adopting a particular

statistical model, as the model makes an assumption about the underlying latent variable structure. As will be discussed in the following section, this dissertation proposes that considering the clinical implications of adopting one over another offers an important point of comparison for deciding whether to analyze clinical data with a dimensional or categorical assumption. Specifically, it is proposed that the clinical implications for decision-making can be considered in model selection. In order to discuss these implications, a general model for clinical decision-making that describes the underlying process by which clinical feedback influences clinical judgment and decisions will first be presented.

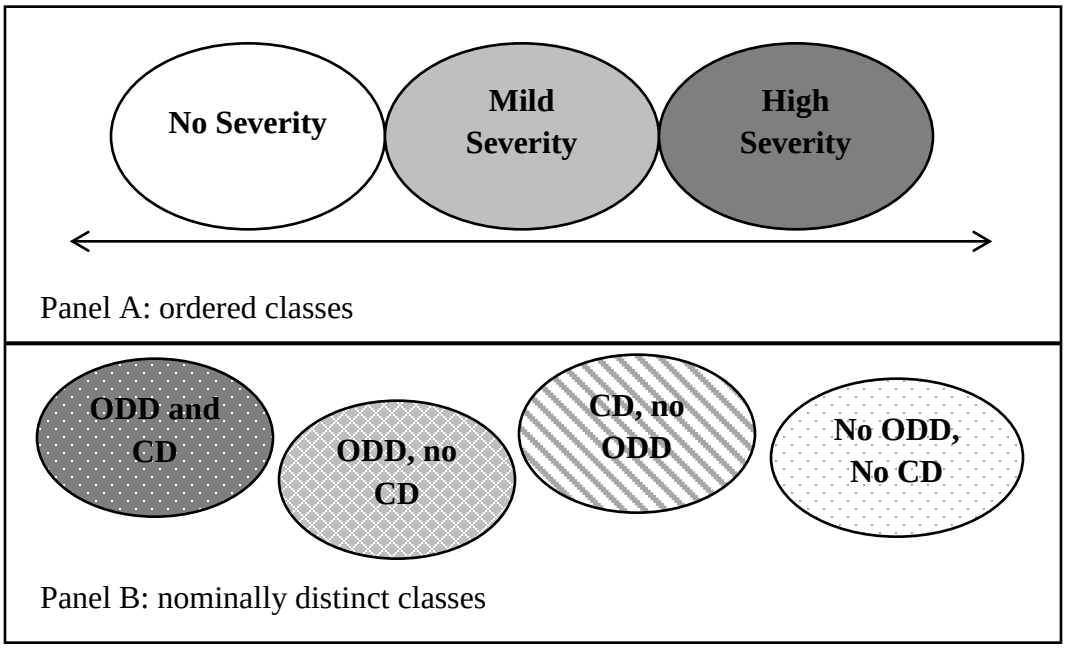


Figure 1. Conceptualization of latent classes representing ordered classes along an underlying continuum (Panel A) or as nominally distinct (Panel B).
Note: ODD = Oppositional Defiant Disorder; CD = Conduct Disorder.

Information Processing and Clinical Decision-Making

Little is known about the underlying mechanism of clinical decision-making. As Hoagwood and Kolko state so eloquently, "it is difficult and perhaps foolhardy to try to improve what you do not understand" (2009, p. 35). The following section proposes a general problem-solving model based on information processing theory to depict the process of clinical decision-making.

Problem-Solving Model

A general model for clinical decision-making can be found in Figure 2. It is based on a general problem-solving model rooted in information processing theory (Vessey, 1991). This basic model identifies the problem solution (here, the clinical judgment, or decision made) as the outcome of the relationship between the problem-solving task (i.e., the type of clinical decision or judgment needed to be made) and the problem representation (i.e., the presentation of information used to inform the problem solution). Here, the word *information* is used to represent the totality of material a clinician has available to work with in making a decision or judgment, and *task* refers to any clinical decision or judgment that the clinician makes in treatment planning. For example, a clinical task may be answering questions such as: Is this client improving? Does this client need more intensive services? Is this therapeutic approach working? etc. In the center of the model is the mental representation, or the way that the task is represented in the clinician's working memory. This mental representation is dependent

on both the task and the information being presented for use in the specific task, and ultimately leads to a solution to the problem, or in the specific case of clinicians, the clinical judgment, or decision made.

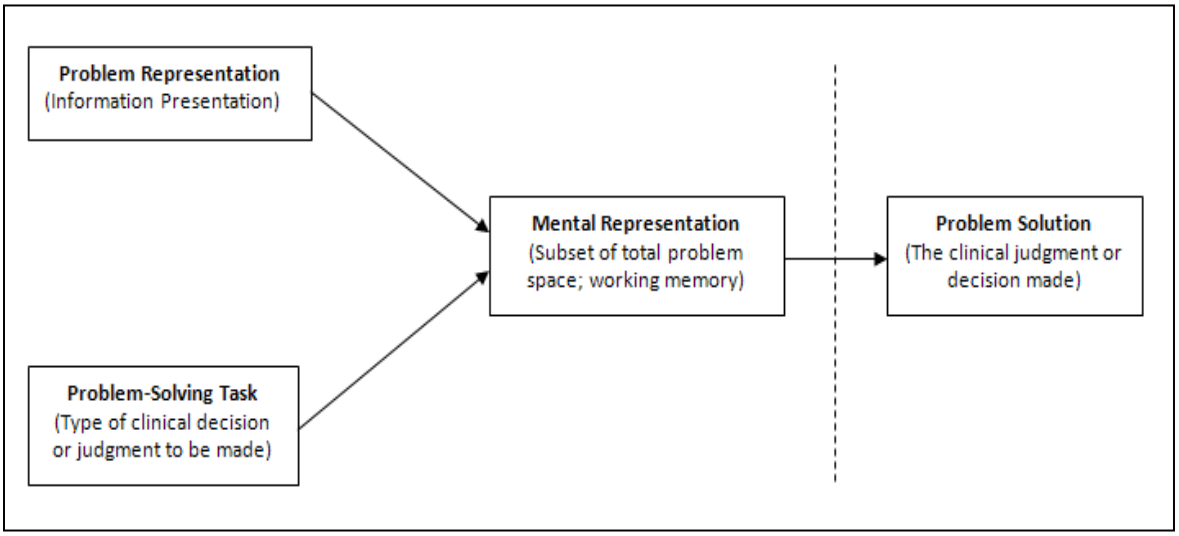


Figure 2. Clinical decision-making represented with a general decision-making model (adapted from Vessey, 1991)

An important point of consideration, however, is that humans have limited information-processing capacity and are therefore unable to process large amounts of information. This is often referred to as *cognitive load*. When the amount of information produces a high cognitive load, individuals seek strategies that lower the cognitive load such as by attending to only a small amount of the information or by utilizing cognitive heuristics to simplify the decision process. These heuristics are cognitive short cuts that aim to oversimplify the problem and therefore may not result in accurate decisions (Kliger & Kudryavtsev, 2010). A few of the most commonly discussed heuristics are briefly described below.

- Availability: the tendency to determine the likelihood of an event according to the easiness of recalling similar or more frequent instances. This can result in assigning more weight to the information most easily recalled instead of processing all the relevant information (Kliger & Kudryavtsev, 2010).
- Representativeness: only specific essential features of an object are examined and compared to the features of a class of objects to see if the object is representative of the class. This may lead to oversimplification or premature assignment to a class prior to integrating all information and features of the object (Benbasat & Taylor, 1982).
- Overconfidence: People tend to have more confidence in their ability to assess probable outcomes than is justified based on evidence. In this case, important information may be ignored or overlooked simply because the person is confident that the outcome is already known (Benbasat & Taylor, 1982).
- Anchoring and adjustment: Anchoring is the natural starting point in any task and becomes the basis for all future judgments. In most cases, adjustment must occur when additional information is provided. However, adjustment in the face of this additional information has often proved to be imprecise and insufficient (Benbasat & Taylor, 1982).

Not only does the amount of information affect the cognitive load, the complexity of the task also asserts influence. Although a common definition of task complexity has not been agreed upon in the information processing literature (Speier, 2006), it often refers to tasks that require information to be interpreted or analyzed (i.e., manipulated

mathematically, assessed and interpreted based on a known norm) or that require non-linear connections or assessment of joint-effects between two information cues. Task complexity is also impacted by the relevance of the information provided, the number of tasks being processed simultaneously, the time pressure to complete the task, and the context in which the task is being performed. In other words, task complexity is highly related to the degree of mental effort required to identify a problem solution (Payne, 1976).

In order to gain a better understanding of what is meant by a complex task, one might consider an example of a 'simple' task presented in the literature. One of the simple tasks utilized by Speier, Vessey, and Valacich (2003) included the presentation of a table including the work capacity (in hours) and workload (in hours) from three work centers over six months. The task involved identifying in which month there was the greatest workload at all three work centers. In this case, participants were required to scan the table and identify a month where the numbers for workload were highest across the work centers. It is important to note that, in this case, there existed a right and wrong answer, all the information needed to make a correct decision was presented, and no external time pressure was exerted. Vessey and Galletta (1991) utilized another example of a simple task where participants were required to respond to bookkeeping questions based on information provided about five bank accounts. Participants were presented with either a graph or a table and asked about the specific number of deposits or withdrawals for a specific month. Again, the task had correct and incorrect answers, all information needed to determine the correct answer was present, and no time pressure

was exerted. These two examples of tasks exemplify 'simple' tasks found in the literature.

In comparison to simple tasks, clinicians face tasks (e.g., decisions and judgments used for case conceptualization, diagnosing, treatment planning, etc.) that require the integration of information from multiple sources concerning unobservable psychopathologies (i.e., only the symptoms are observable). Additionally, with high caseloads and little time between clinical sessions, clinicians often experience time pressure. All of these things have the potential to increase the cognitive load. Based on this, it is safe to assert that clinical judgments and decisions represent complex tasks. Therefore, attention to the cognitive load of clinicians is particularly important in understanding how feedback influences clinical judgment and decision-making. If the clinician is faced with too much information, or irrelevant or otherwise inappropriate information, his/her decision-making ability may be compromised.

Although previously used to reflect the totality of information available for a judgment or decision, the construct of *information* is now narrowed to reflect only the data provided to clinicians in the form of clinical feedback. As described above, evidence consistently supports that the provision of clinical feedback to clinicians improves client outcomes (e.g., Bickman et al., 2011; Lambert et al., 2005). In psychotherapy research, the most common form of clinician feedback includes indicators of patient progress (e.g., symptom severity) that are based on standardized outcome measures. It is a commonly held belief among researchers that the use of such clinical feedback provides ongoing information concerning client progress that can aide clinicians in making treatment decisions (e.g., Fishman, 2001; Stein, Kogan, Hutchison, Magee, &

Sorbero, 2010). Despite this, review of clinical literature yielded no empirical research investigating how best to present client information in feedback. This presentation may matter. For example, within the decision-making and information processing literature, it is widely recognized that the presentation of information (as a table versus a graph, for example) can significantly affect decision-making (Hwang, 1994; Speier, 2006; Speier et al., 2003; Vessey, 1991). Therefore, attention to how client information is presented as feedback to clinicians is important to determine whether it affects clinical decision-making. In the next section, the theory of cognitive fit is introduced and applied to the presentation of information in feedback used for clinical decision-making.

Cognitive Fit

The concept of cognitive fit provides a theoretical basis for understanding how information presentations support decision-making tasks (Vessey, 1991). According to this concept, task performance is enhanced when there is a match or fit between the information that is emphasized in the information presentation (Problem Representation in Figure 2) and the information that is needed to solve the particular task at hand (Problem-Solving Task in Figure 2). This occurs because the decision-maker is easily able to understand and interpret the information directly as it is needed for application to the specific task, thus reducing cognitive load. Alternatively, by definition, task performance accuracy is decreased when there is a mismatch between the information presentation and the task. In this case, the lack of cognitive fit increases the cognitive load by requiring the individual to transform the information presented into a form that is relevant to the task being performed.

Most research on cognitive fit has investigated information presentation and tasks defined as either spatial or symbolic (Hwang, 1994; Speier et al., 2003; Vessey, 1994). Information presented in tables is considered symbolic where discrete sets of symbols (or data) are presented (e.g., names, orders of quantities, table of airplane arrival/departure times). On the other hand, information presented in graphs is considered spatial (e.g., change in amount over time, relationship among two different performances, comparison of sales regions). Based on cognitive fit theory, decision-making accuracy is maximized when data are presented in a table (i.e., symbolically) for symbolic tasks (e.g., recall of specific values), and when data are presented in a graph (i.e., spatially) for spatial tasks (e.g., assessment of trends or relative increase or decrease in value).

As was previously discussed, data collected on standardized outcomes measures can be analyzed based on the assumption that the underlying latent variable is categorical or dimensional in structure. This assumption changes the nature of the information that is presented as feedback. Data analyzed assuming a categorical latent structure will result in feedback reflecting this categorical assumption (i.e., differences in groups). Similarly, data analyzed assuming a dimensional latent structure will result in feedback reflecting this dimensional assumption (i.e., individual differences by degree). Application of the concept of cognitive fit would propose feedback presenting categorical information is ideal for categorical tasks and that feedback presenting dimensional information is ideal for dimensional tasks. Under these matched conditions, the cognitive load is minimized such that information can be directly applied to the task, resulting in efficient and effective decision-making (see Figure 3). In this way, investigating how the presentation

of information matches with clinical tasks is important for understanding the role of cognitive fit for clinical decision-making.

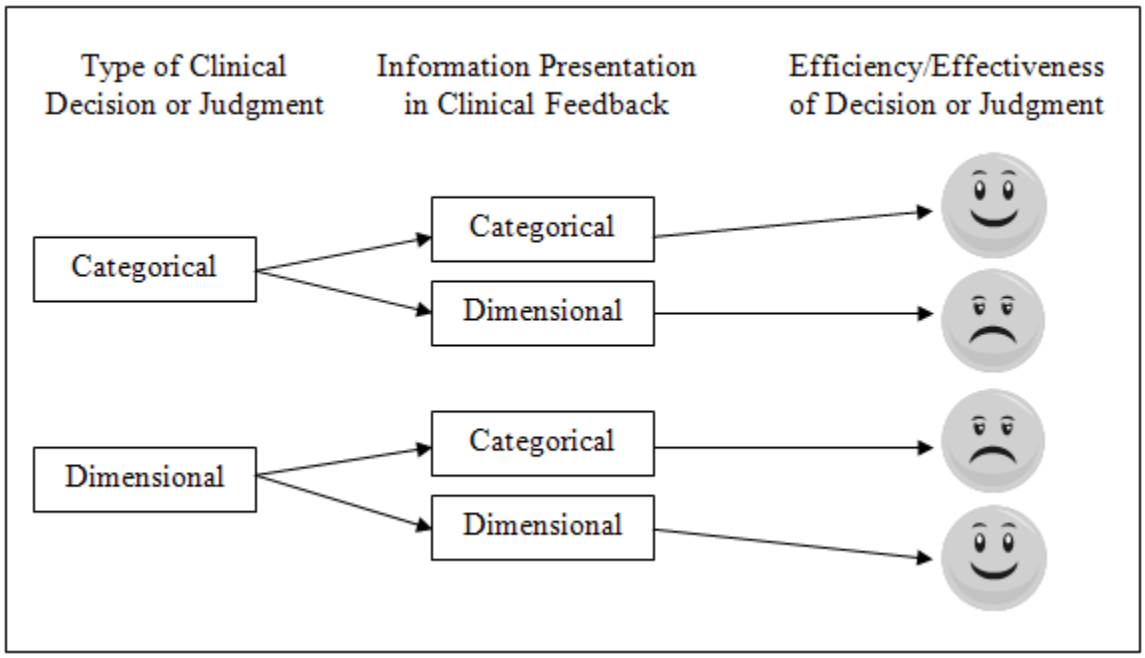


Figure 3. Cognitive fit as applied to clinical decision-making with clinical feedback

The following section connects the application of cognitive fit as related to clinical decision-making to the ongoing debate concerning whether latent clinical variables are assumed to be categorical or dimensional in nature. It is proposed that the latent variable structure that produces feedback presented in a format matching the clinical task results in more efficient and effective clinical judgment and decisions compared to a format that does not match. This section will also introduce a pilot clinical survey to gather insight from trained clinicians concerning what presentation of information (dimensional or categorical) they perceive as being most applicable to specific clinical tasks.

Using Cognitive Fit as Criteria for Model Selection

Representing latent clinical variables as either categorical or dimensional has advantages and disadvantages, as well as proponents and opponents. However, the statistical model used to analyze data assumes a particular structure and produces output that presents the clinical information based on this assumption. For example, data analyzed with IRT is assumed to have a dimensional latent variable structure, resulting in output reflecting a dimensional structure. Individuals are then described based on their individual differences in the degree of the latent variable. Similarly, data analyzed with LCA is assumed to have a categorical latent variable structure, resulting in output reflecting a categorical structure. Here, groups are differentiated from one another. The output produced by either analytic approach is used for clinical feedback meant to provide clinicians with information about specific client status and progress. In other words, clinicians may either receive information about a client's degree of externalizing severity (dimensional) or about their membership to a distinct latent group that describes their severity (categorical). Although there may be overlap between the overall nature of the information provided from these two approaches, it is important to consider the effect these different information presentations have on clinical practice. This is conceptualized via the concept of cognitive fit.

Based on cognitive fit, some information presentations are more effective for decision-making depending on the task the information is being used to inform. When there is a match between the information presentation and task, cognitive load is minimized and decision-making is more efficient and effective. When there is not a

match, more cognitive effort is needed to transform the information into a usable form for the specific task, thus increasing cognitive load and decreasing decision-making efficiency and effectiveness. In this sense, one latent structure may not be *better* than another in a pure sense, but rather better in terms of producing a presentation of clinical information that maximizes cognitive fit, thus supporting effective and efficient clinical decision-making. Because tasks can be different in nature and require different cognitive processes, it may be that one presentation is matched with one specific task and another presentation is matched with another task. Therefore, in applying the theory of cognitive fit to clinical decision-making, it is important to describe and classify the nature of the specific tasks found within the practice of clinical psychology. Only then can tasks be hypothesized to be matched or mismatched with specific information presentations.

Clinicians make a wide range of clinical decisions and judgments throughout the process of evaluating and treating clients; some of the common clinical tasks (i.e., judgments or decisions) are listed in Table 1 below. These were collected and inferred from previous literature on clinical judgment and are listed as questions that represent the decisions or judgments clinicians may make for treatment planning. While not meant to be exhaustive of the decisions and clinical judgments made by clinicians on a day-to-day basis, this list illustrates how broadly tasks can vary. Some appear to involve more black-or-white 'yes' or 'no' decisions (e.g., to hospitalize or not, to diagnose or not); others reflect a more subtle judgment that can range from a definitive 'yes' to a definitive 'no' and every shade of gray in between (e.g., determining whether the client getting better or not, whether the treatment is working or not). In either case, the clinician's answers to those questions will guide the client's treatment and ultimately affect

outcomes. Therefore, it is important that information provided to clinicians as an aid for making each decision or judgment is presented in such a manner that facilitates the most effective and efficient decision-making.

Table 1. Common Clinical Judgments and Decisions (Phrased as Questions)

- Does this client display clinical levels of psychopathology?
- What level of care is needed for this client?
- Should I assess (or reassess) this client for a diagnosis?
- Should this client be referred for medication consultation?
- How frequent does this client need to have sessions?
- Is the treatment working for this client?
- Is this client getting better?
- Is this client developing new symptoms?
- Is this client deteriorating?
- Does this client require hospitalization?
- Should the client's treatment plan be reviewed or changed?
- Is this client ready for termination of treatment?

As mentioned earlier, previous work in other fields investigating cognitive fit have clearly labeled tasks as *spatial* or *symbolic*. However, these were simple tasks such as determining a trend from a graph or deciding whether an object was the same or different from another. Clinical tasks are more complex and uniquely understood by those who have been clinically trained. Therefore, instead of naïvely proposing that a particular task matches with a presentation of information, feedback was collected from colleagues who are trained and/or practicing clinicians concerning the presentation of information they find most conducive to making specific clinical decisions and judgments. Clinicians were given general examples of both categorical and dimensional representations of symptom severity information and asked to respond, on a 5-point

Likert-type scale (from 1 = *not useful* to 5 = *very useful*), how useful each presentation would be to inform the specific clinical tasks listed in Table 1. A copy of this informal survey can be found in Appendix A. The information gained from this survey (to be presented in chapter 4) provides a foundation for illustrating the potential role of cognitive fit without assuming any knowledge of the nature and intricacies of the clinical tasks faced by clinicians throughout their daily practice. Clinicians may already be naturally aware of which presentations are matched with specific clinical decisions or judgments. In other words, clinicians may choose the presentation that shows the best cognitive fit to each clinical task. Drawing on their insight provides a basis for initially proposing what presentations and tasks may be matched and also provides a starting point for testing the theory of cognitive fit in regards to the use of feedback in clinical decision-making.

If the theory of cognitive fit holds with the use of feedback in clinical decision-making, then there exists an ideal match between the presentation of information provided in the feedback and the specific clinical task. When this ideal match is made, the clinician's cognitive load is minimized and the clinician is able to make more effective and efficient decisions. This ultimately could lead to improved clinical outcomes, a universal goal of researchers and practitioners alike. Because the presentation of clinical information is highly dependent on the structural assumptions of the latent construct made by the specific statistical model applied, understanding this match is particularly salient for deciding whether to use a model assuming a categorical or dimensional latent variable structure. For example, a categorical approach is most appropriate for a decision that relates to a group differences whereas a dimensional

approach is most appropriate for a decision that relates to individual differences in degree. In this way, exploring cognitive fit provides a practical and potentially impactful way of comparing these different structural conceptualizations.

Summary

So far, this dissertation: 1) highlights the importance of clinical decision-making in mental health treatment; 2) briefly summarizes the limited research investigating clinical decision-making; and 3) discusses the effectiveness of providing clinical feedback to improve clinical outcomes. Next, this dissertation: 4) discusses the assumptions made about the underlying structure of latent clinical variables; 5) presents specific issues in clinical measurement based on these assumptions; and 6) presents research based on these assumptions as applied to the specific clinical construct of externalizing symptom severity. Then, this dissertation: 7) introduces a general model of decision-making and applies it to clinical judgment and decision-making; 8) adopts the concept of cognitive fit whereby the presentation of information is matched with the problem-solving task to allow more effective and efficient decision-making (aim 1); and 9) applies cognitive fit to specific information presentations in clinical feedback. Finally, this dissertation: 10) proposes the use of cognitive fit for evaluating the use of different latent variable structures; 11) lists specific clinical decisions and judgments and identifies the need to examine what information presentation formats are most conducive to their successful completion; and 12) briefly describes the method for matching information formats to tasks by utilizing feedback gathered from trained clinicians.

Chapter 2 – 4 will extend and further illustrate the application of cognitive fit to clinical decision-making with an empirical application using clinical data from youths aged 11 -18 receiving home-based mental health treatment. Specifically, data were gathered from the caregiver version of the externalizing subscale of the Symptoms and Functioning Severity Scale (SFSS). Statistical models assuming a categorical and a dimensional latent variable structure will be applied to the data both at a single time point (i.e., cross-sectional) and across two time points (i.e., longitudinal). Thus, four models will be fit to the data (see Table 2), each producing model output that is dependent on the structural assumption for the underlying latent variable: Model 1 is a latent class analysis (LCA) model reflecting a cross-sectional categorical approach; Model 2 is a graded response model (GRM) reflecting a cross-sectional dimensional approach; Model 3 is a latent transition analysis (LTA) model reflecting a longitudinal categorical approach; and Model 4 is a longitudinal GRM reflecting a longitudinal dimensional approach.

Based on the assumptions of the four models in Table 2, the presentation of information will differ based on how the latent clinical variable is modeled structurally (e.g., as categorical versus dimensional). The goal of this empirical application is twofold. First, the outputs from these statistical models will be compared across individuals and groups. This serves as aim two of the overall dissertation. While there are an increasing number of articles comparing the statistical fit of models with different latent structures (for example, see Bauer & McNaughton Reyes, 2010; Hartman et al., 2001; Walton et al., 2011), none have directly compared the model output from models assuming each different structure. Given that, in the clinical context, decisions are made based on the individual client information (Lutz, Böhnke, & Köck, 2011), illustrating the

similarities and differences across models for the *same* client is particularly salient. Such comparisons will not only allow for a better understanding of the consequences for an individual client when selecting a particular statistical model, but the use of real-world clinical data will provide a clearer illustration of the potential effect of selecting a categorical or dimensional approach on clinical decision-making.

Table 2. Description of Four Proposed Models

Time	Construct Form	Model	Model Description
Cross-Sectional	Categorical	1	Latent Class Analysis (LCA) Model
	Dimensional	2	Graded Response Model (GRM)
Longitudinal	Categorical	3	Latent Transition Analysis (LTA) Model
	Dimensional	4	Longitudinal GRM

The second goal of the empirical application is to select individual clients from the sample and, using outputs from each of the four models, expand the illustration of cognitive fit by discussing the specific types of clinical tasks that may be more or less matched with the different presentations of clinical information as would appear in feedback. This is aim three of the overall dissertation. This discussion will utilize information gathered from trained clinicians concerning their data presentation preferences (i.e., categorical or dimensional) for specific clinical tasks. This will help examine the concept of cognitive fit within the clinical context by highlighting the

potential impact on clinical decision-making for real clients represented in the data. The result from this will lead to some proposals for future research that will further explore and evaluate the potential role of cognitive fit for determining how to present information in clinical feedback. This serves as aim for of the overall dissertation.

In conclusion, the four major aims of this dissertation are intended to illustrate the concept of cognitive fit and propose how this concept may be used to determine how to conceptualize the structure of latent clinical variables when statistically analyzed for presentation in clinical feedback. It is asserted that the structural conceptualization that should be used is the one that produces clinical feedback that presents clinical information in a way that is most conducive for efficient and effective decision-making. In this way, the ultimate support for selecting one latent structure over another is dependent on how its application has the potential to affect clinical outcomes.

The next chapter (Chapter 2) presents the sample, measures, and specific methods utilized in the empirical application. This chapter also includes model descriptions of the four statistical models listed in Table 2 that are the basis for the empirical application used to accomplish aim two. Chapter 3 presents a comprehensive psychometric analysis of the SFSS Externalizing subscale, as well as testing for the assumption of unidimensionality that is essential for IRT modeling. This basis for this will be discussed further in Chapter 2. Chapter 4 presents the results of fitting the statistical models to the data, as well as the results of the informal clinician surveys used in aim three. Finally, Chapter 5 is the discussion, which includes aim four: proposals for future research.

CHAPTER 2

Empirical Application Methods

Sample and Measures

This chapter describes the sample and data used in the empirical application (aim two). Additionally, this chapter presents the model equations for the four models applied to the data (refer to Table 2) and the respective statistical procedures for their application with Mplus version 6.11 (Muthén & Muthén, 2010).

Participants

Participants were from a larger study evaluating the effects of a measurement feedback system (Contextualized Feedback Systems; CFS™) on youth outcomes (Bickman et al., 2011). This sample was drawn from 28 regional offices in 10 different states comprising part of a large national provider for home-based mental health services. This service provider is a highly decentralized organization where the specific type of treatment is not prescribed. Therefore, treatment could include individual and family in-home counseling, intensive in-home services, crisis intervention, life skills training, substance abuse treatment, and case management. Clinicians utilized various therapeutic techniques including cognitive-behavioral, integrative-eclectic, behavioral, family systems, and play therapy.

The sample selection process for the current paper is depicted in Figure 4. The sample for the current paper included all youth in the larger evaluation study of CFS™ (Bickman et al., 2011). These were youth who began treatment during the two-and-a-half year data collection period (*N* = 356). Therefore, the first data point for each youth was regarded as the beginning of treatment (time 1). Additionally, inclusion in current analyses required having at least two valid caregiver SFSS Externalizing Subscale measures. A valid measure was defined as completion of at least 80% of SFSS Externalizing items. This resulted in a final sample of 204 youth receiving mental health treatment, and their respective caregivers and clinicians. The last time point in which caregiver SFSS Externalizing data was present for each youth was used as the final (time 2) measurement.

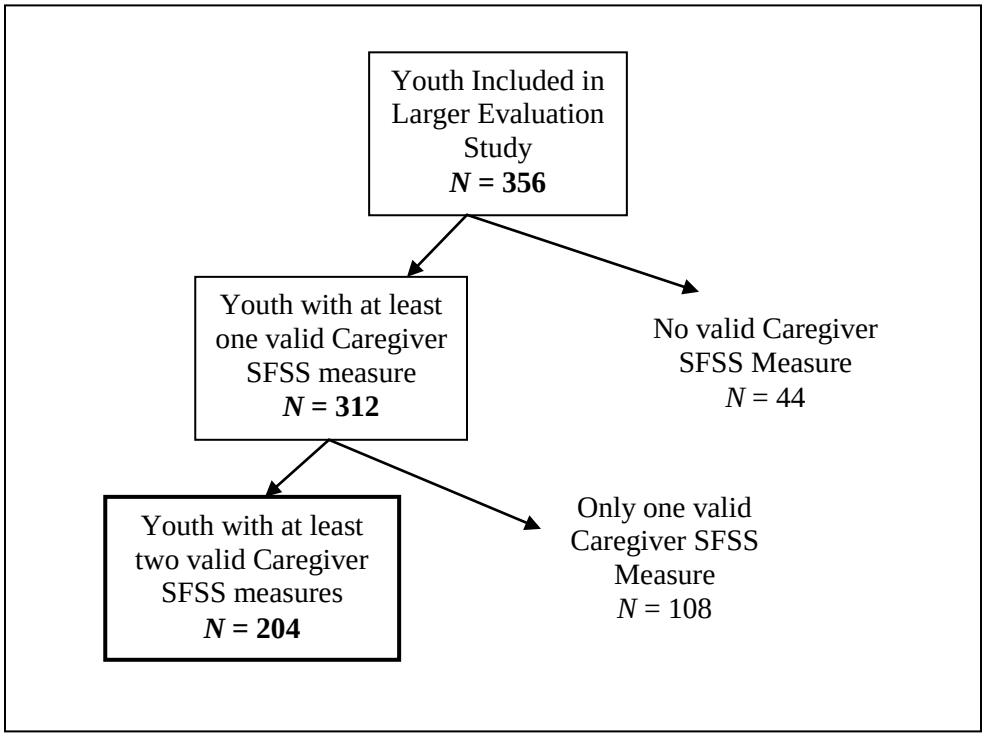


Figure 4. Sample selection process for current study

In order to test for potential selection bias, several comparisons were made. First, baseline characteristics of youth and caregivers with at least one valid caregiver SFSS Externalizing subscale (N = 312) were compared to those without one (N = 44). There were no statistically significant differences between these groups based on youth age, gender, racial background, or caregiver report of externalizing symptom severity. Additionally, no significant differences existed between these two groups in regards to caregiver age, gender, highest level of education, racial background, marital status, household income, or relationship to the youth. The only aspect that distinguished these groups was those without a valid caregiver SFSS Externalizing subscale remained in CFS™ a significantly shorter amount of time (M = 6.94 weeks, SD = 1.20) compared to those with a valid measure (M = 17.11, SD = 13.82; $t(354) = -4.76$, $p < 0.001$).

Next, baseline characteristics of youth and caregivers with at least two valid caregiver SFSS Externalizing subscales (N = 204) were compared to those with only one (N = 108). Only two statistically significant differences were detected. First, youth whose caregiver completed at least two SFSS Externalizing subscales were significantly younger (M = 14.56 years, SD = 1.78) than those whose caregivers completed only one (M = 15.20 years, SD = 1.85; $t(312) = -3.01$, $p < 0.01$). Second, youth whose caregivers completed the measure at least twice were in CFS™ significantly longer (M = 21.24 weeks, SD = 13.79) compared with youth whose caregiver completed only one SFSS Externalizing subscale (M = 9.30 weeks, SD = 10.01; $t(310) = 7.96$, $p < 0.001$). There were no significant differences between these groups based on youth gender, racial background, or age. Similarly, there were no significant differences based on caregiver age, gender, level of education, marital status, or relationship to the youth. Additionally,

the baseline caregiver SFSS scores (total and subscales) did not differ significantly between these groups. Thus, the number of completed caregiver SFSS Externalizing subscale does not appear to be a function of caregiver-rated baseline youth symptom severity.

Youth included in this study (N = 204) ranged in age from 11 – 18 years (M = 14.56, SD = 1.77) and 48% were female. Caregivers of these youth ranged in age from 23 to 77 years (M = 43.43, SD = 10.62) and 87.7% were female. Nearly all caregivers indicated they were the primary caregivers for the youth (96.5%). Breakdowns of several other youth and caregiver background variables can be found in Table 3. For longitudinal models, the first and last measurement points with a valid caregiver SFSS measure (time 1 and time 2 respectively) were used to represent baseline and final measurement of youth externalizing symptom severity. Due to the real-world nature of the larger evaluation study, youth remained in the study (i.e., received treatment) for differing lengths of time. Therefore, baseline and final measurement points reflect the unique length of time each youth received treatment. Time between baseline and final measurement points in the current longitudinal study ranged from 2 to 62 weeks (M = 16.36, SD = 12.26). Although youths were in treatment for considerably different lengths of time, this may better reflect 'typical' treatment length often found in community mental health counseling. This is in contrast to laboratory studies where researchers assign all participants to receive a fixed amount of treatment (ex. three or six months).

Table 3. Demographics of Youth and Caregivers in Analytic Sample (*N* = 204)

	N (valid %)
Youth Characteristics	
Racial/Ethnic Background (missing *n* = 22)	
Caucasian	108 (59.3)
African American	40 (22.0)
More than one race	23 (12.6)
Other	11 (6.1)
Caregiver Characteristics	
Racial/Ethnic Background (missing *n* = 55)	
Caucasian	121 (68.4)
African American	45 (25.4)
More than one race	2 (1.1)
Other	9 (5.1)
Marital Status (missing *n* = 53)	
Married/Living as Married	77 (43.5)
Divorced/Separated	61 (34.5)
Never Married	29 (16.4)
Widowed	10 (5.6)
Highest Level of Education (missing *n* = 60)	
Less than a High School Diploma/GED	40 (23.7)
High School Diploma/GED	99 (58.6)
College/post-grad Degree	30 (17.8)
Household Income (missing *n* = 74)	
Less than $10,000	38 (25.0)
$10,000 - $19,999	32 (21.1)
$20,000 – $34,999	40 (26.3)
$35,000 - $49,999	20 (13.2)
$50,000 or more	22 (14.5)
Relationship to the youth (missing *n* = 54)	
Birth parent	122 (68.9)
Adoptive parent	10 (5.6)
Grandparent	19 (10.7)
Step-parent	1 (0.6)
Family member	14 (7.9)
Foster parent	7 (4.0)
Other	4 (2.3)

Measures

Symptoms and Functioning Severity Scale (SFSS). The SFSS was designed for use with youth aged 11 to 18 years old receiving mental health services (Bickman et al., 2007). The SFSS provides a global measure of the youth's symptom severity and functioning as well as subscale scores for the severity of externalizing and internalizing emotional and behavioral problems. Composed of 33 items, the original version (v.1) of the SFSS is completed by three respondents: the youth, the youth's parent or primary caretaker, and the clinician. The measure takes 5 – 7 minutes to complete at the end of a clinical session, is written at a fourth grade reading level, and is designed to measure change over time in closely repeated measurements (e.g., every other session). Modeled after the CBCL, items chosen for SFSS represent symptoms/behaviors associated with the most common mental health disorders for youth: ADHD, CD, ODD, depression, and anxiety. The SFSS has demonstrated sound psychometric qualities for all respondent forms including internal consistency (Cronbach alpha range = .93 –.95), test-retest reliability (*r* range = .68 –.87) and convergent and discriminant validity. Additionally, confirmatory factor analysis (CFA) has confirmed the two correlated factors (externalizing and internalizing) structure for all three respondent forms of the SFSS. For more information about the psychometric qualities of the SFSS measure, see the first edition of the Peabody Treatment Progress Battery manual (PTPB; Bickman et al., 2007). A copy of the full SFSS is found in Appendix B.

The current study utilized the caregiver version of the SFSS Externalizing subscale only. The SFSS Externalizing subscale is composed of 16 Likert-type questions rated 1 to 5 (*never, rarely, sometimes, often, very often*). For the purposes of the present

study, response categories were collapsed into three categories (*never, rarely/sometimes, often/very often*). This was done because low frequency of category endorsement can cause problems with the estimation of IRT parameter estimates (see for example Eid & Diener, 2001; Gadermann, Schonert-Reichl & Zumbo, 2010; Vitterso, Biswas-Diener & Diener, 2005 for analogous strategies). Chapter 3 presents a comprehensive psychometric analysis of the SFSS Externalizing subscale with these reduced three response categories.

Youth and caregiver background forms. As part of the larger evaluation study of CFS™, caregivers and youth completed a background form during their initial/baseline session. This form included items about caregiver and youth background profiles such as age, gender, relationship, and previous diagnoses.

Clinician initial assessment form. As part of the clinical intake within CFS™, clinicians completed an initial assessment form. This form included information such as the youth's presenting problems, previous mental health evaluations or school services received, use of alcohol/drugs, presence of previous DSM-IV psychiatric diagnosis, the inclusion of a DSM-IV psychiatric diagnosis with the current intake, as well as the youth's most recent primary and secondary DSM-IV diagnoses (if any) .

Procedures

Caregivers completed the SFSS at the end of the clinical session according to the measurement schedule included as part of the larger evaluation study. This schedule recommended completion of the SFSS at treatment baseline and every other week throughout treatment. Completed measures were sealed in an envelope and

administrative assistants at each clinical site entered data into the CFS™ system. Therefore, only de-identified data were received for the current analyses. The Institutional Review Board of Vanderbilt University granted approval for the research design and the data collection procedures of the CFS™ evaluation study.

Statistical Analyses

Four models were fit in the current dissertation. The first two are cross-sectional models: 1) the latent class analysis (LCA) model; and 2) the graded response model (GRM). The remaining two are longitudinal extensions of these first two models: 3) the latent transition analysis (LTA) model; and 4) the longitudinal GRM. The LCA and LTA models are latent class models that assume the latent variable is categorical whereas the GRM and longitudinal GRM are item response models that assume the latent variable is continuous. All analyses in the current dissertation were conducted with Mplus version 6.11 (Muthén & Muthén, 2010) that utilizes a maximum likelihood estimation with robust standard errors. Accordingly, these four models are specified with Mplus formulation in the following section. Annotated Mplus syntax for sample models used in the current dissertation can be found in Appendix C.

The SFSS Externalizing subscale is composed of 16 Likert-type items with 3 ordered response categories coded 0, 1 or 2. In the current application, participants completed the subscale at two time points during treatment. Although approximately 90% of the participants included in these analyses had complete data (i.e., they answered each at each time point), some data were missing. Therefore, following procedures

suggested by McKnight, McKnight, Sidani, and Figueredo (2007), patterns of missingness for item responses were inspected across subjects and items. Because no discernible patterns of missingness were found that indicated missing data were not missing at random (non-MAR), all available data were used under the missing-at-random (MAR) assumption as defined by Little and Rubin (1987). Mplus allows for this missing data in using full information maximum likelihood estimation (FIML; for more information on FIML, see Enders & Bandalos, 2001).

Definition of Subscripts, Parameters, and Symbols

For consistency, the model equations presented in this chapter will utilize the same naming conventions for subscripts, parameters, and symbols (when applicable). These are as follows:

Subscripts:

i is an indicator of an item where $i = 1, ..., I$,

j is an indicator of a person where $j = 1, ..., J$.

k is an indicator of an item category where $k = 1, ..., m$,

g is an indicator of a latent class where $g = 1, \ldots, G$,

and

t is an indicator of time where $t = 1, ..., T$.

Parameters:

θ is a trait score (i.e., level of externalizing symptom severity),

δ is an item threshold,

and

λ is a factor loading.

Symbols:

Y is a random variable representing the response,

Y^* is a continuous latent response variable,

and

y_k is an observed category score (i.e., a realization of the random variable) corresponding to category k.

Model 1: Latent Class Analysis Model

Latent class analysis (LCA; Clogg, 1995; Goodman, 1974; Lazarsfeld & Henry, 1968) is a probabilistic statistical technique used for detecting unobserved homogeneous subgroups ("latent classes") of individuals based on their responses to test items. LCA assumes that a number of discrete latent classes explain all individual difference and that no associations among items exist within class (i.e., axiom of local independence). In this way, the latent classes or discrete latent variable fully accounts for associations between the observed item responses. Therefore, each latent class is characterized by a pattern of conditional probabilities that indicate the probability of certain item responses (Collins & Lanza, 2010). The probability of person j having Y_{ji} higher than or equal to y_k on item i given their latent class membership is:

$$\Pr(Y_{ji} \geq y_k \mid C_j = g) = \frac{\exp(\delta_{ikg} - Y_{ji}^*)}{1 + \exp(\delta_{ikg} - Y_{ji}^*)}, \quad (1)$$

where C_j is a discrete latent variable denoting latent class membership and g is a specific individual-level discrete latent variable. δ_{ikg} is an item- and latent class-specific threshold parameter where δ_{i1g} is the threshold between response category 0 and 1 & 2 for item i in latent class g, and δ_{i2g} is the threshold between response category 0 & 1 and 2 for item i in latent class g, when there are three response categories. Y^*_{ji} is a continuous latent response variable for person j on item i such that:

$y_k = 0$ if $Y^*_{ji} \leq \delta_{i1g}$,

$y_k = 1$ if $\delta_{i1g} < Y^*_{ji} \leq \delta_{i2g}$,

and

$y_k = 2$ if $Y^*_{ji} > \delta_{i2g}$.

This relationship between the continuous latent response variable, the item thresholds, and the observed category scores is depicted in Figure 5.

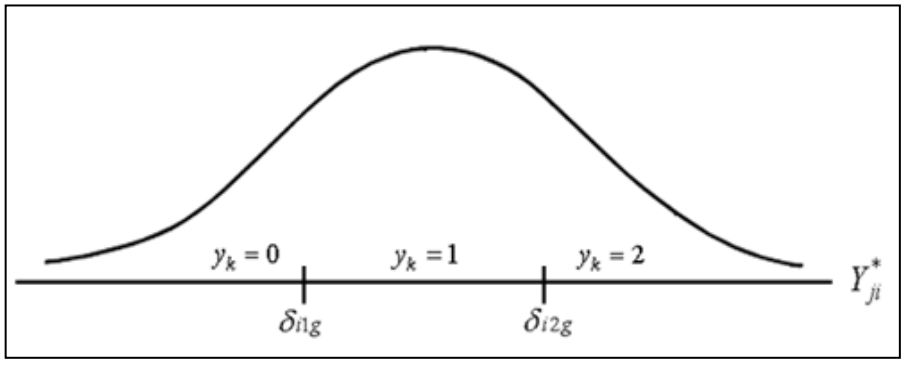


Figure 5. Relationship between continuous latent response variable (Y^*_{ji}), item thresholds (δ_{ikg}), and observed category scores (y_k).

The goal of LCA is to identify the smallest number of mutually exclusive latent classes that adequately describe the associations among the observed item responses. Determining the smallest number of classes is accomplished by comparing models with different numbers of latent classes and selecting a final model based on predefined selection criteria. Although numerous model selection criteria can be used to compare models with differing numbers of latent classes (Magidson & Vermunt, 2000; McCutcheon, 1987), there is no consensus on the best criteria. The current dissertation used the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Lo Mendell Rubin Likelihood Ratio Test (LMR LRT; Lo, Mendell, & Rubin, 2001). The BIC is commonly used and has been examined extensively in the context of latent variable modeling (e.g., Celeux & Soromenho, 1996; Fishler, Grossman, & Messer, 2002; Yang, 1998). In simulation studies, the BIC has performed well in terms of consistency for identifying the 'true' population model (Nylund, Asparouhov, & Muthén, 2007).

The LMR LRT, a more recently proposed criterion index, uses a robust maximum likelihood estimator to test the significance of a difference between the -2 log-likelihoods for a model with G and $G-1$ classes. The LMR LRT treats these models as nested, where a more restricted model with G class is obtained from a less restricted $G-1$ class model by setting the probability of latent class membership in one of the classes at zero. While it is well known that restricting a parameter value on the border of admissible parameter space results in a likelihood ratio that does not follow a chi-square distribution (McLachlan & Peel, 2000), the LMR LRT still utilizes this likelihood but first derives its correct distribution. Therefore, the LMR LRT approach for testing nested models has been recommended for determining the optimal number of latent classes, as

opposed to likelihood ratio tests that tend to overestimate the number of latent classes (Nylund et al., 2007).

In the current dissertation, selecting the final model proceeded in several steps. First a 1-class model was fit to the data. Then additional models were fit sequentially, each time adding one more latent class. These models were then compared and the number of latent classes was determined based on the best fitting model according to the BIC and LMR-LRT indices.

Mplus uses the Expectation-Maximization Algorithm (EM Algorithm; Dempster, Laird, & Rubin, 1977) for parameter estimation of the LCA model. The EM algorithm maximizes the likelihood function of the complete data (i.e., the observed and latent variables) by utilizing two steps: the expectation (E) step and the maximization (M) step. The E-step estimates the expected frequencies of the complete data under the model assumptions and preliminary parameter values conditional on the observed data. The M-Step computes new parameter estimates maximizing the likelihood of this complete data. The E- and M- step create an iterative sequence until convergence, yielding maximum likelihood estimates of the model parameters. Based on these estimated parameters, Bayes' theorem can be used to compute each individual's posterior probability of membership in each latent class based on their pattern of item responses. In the current analyses, participants were assigned to latent classes based on their largest posterior probability.

One potential problem associated with the EM algorithm for estimating LCA model parameters is that of obtaining a local solution (Titterington, Smith, & Makov, 1985). This happens when an estimation algorithm converges on a local maximum

instead of the globally best solution (Wu, 1983). A recommended method for checking whether a local solution was obtained is to run a model with different random starting values to verify the same solution is reached each time (McLachlan & Peel, 2000). While not eliminating the possibility of obtaining a local solution, if the same solution is reached from different starting values, there is increased confidence that it is the global solution. Therefore, for each LCA model in the current application, 200 random start values were drawn and the best 20 optimizations were used. A solution was selected if the final log-likelihood value was replicated multiple times.

Mplus results from LCA in place arbitrarily labels to latent classes such as *Latent Class 1*, *Latent Class 2*, etc. Therefore, once the final model was selected, the class-specific item thresholds as well as the mixture proportions were inspected in order to rename each latent class based on an appropriate clinical interpretation (i.e., *Clinical Severity Class*, *Subclinical Severity Class*, etc.). The arbitrary labeling of latent classes in Mplus results leads to a potential problem called "label switching." This may occur because even if the same model is run on the same data set, the latent class labeled as Latent Class 1 in the results from the first run may be labeled as Latent Class 2 during the second run. Because of this, inspection of item thresholds and mixture proportions is important to ensure that interpretations of latent classes are appropriate and consistent.

Model 2: Graded Response Model

The GRM (Samejima, 1969, 1997) is a two-parameter IRT model frequently used for clinical measurement when items have ordered response options (Aggen, Neale, & Kendler, 2005; Cole et al., 2011; Cooper & Gomez, 2008; Normand, Belanger, & Eisen,

2006; Reise & Waller, 2009;). As specified in Mplus, the probability of person j having Y_{ji} higher than or equal to y_k on item i is defined as:

$$\Pr(Y_{ji} \geq y_k \mid \theta_j) = \frac{\exp(-\delta_{ik} + \lambda_i \theta_j)}{1 + \exp(-\delta_{ik} + \lambda_i \theta_j)}, \quad (2)$$

where δ_{ik} indicates the item thresholds between the response categories, λ_i is the factor loading for item i, and θ_j is the trait level for person j [1]. Trait levels are assumed to be normally distributed with a mean of 0 and variance of 1.

Before applying IRT, it is important it evaluate two important assumptions underlying these models: unidimensionality and local independence. While the unidimensionality assumption implies that the scale items are measuring only one latent trait, the assumption of local independence implies that no relationship is present between a person's responses to different items after accounting for his/her latent trait level (Lord, 1980). Given that unidimensionality is a sufficient condition for satisfying the assumption of local independence, the current dissertation assesses unidimensionality with exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to determine whether these IRT assumptions are met. The results from these analyses are presented as part of the comprehensive psychometric evaluation of the SFSS Externalizing subscale found in Chapter 3.

[1] Mplus formulation differs from traditional IRT formulation. To convert Mplus parameters into traditional IRT item parameters (*a* and *b* for an item threshold and an item discrimination respectively) the following formulas are used: $a = \lambda\sqrt{\psi} / D$ and $b = (\delta - \lambda\alpha) / \lambda\sqrt{\psi}$, where α is the factor mean, ψ is the factor variance, and D = 1.7.

Model 3: Latent Transition Analysis Model

Latent Transition Analysis (LTA; Collins & Wugalter, 1992) is a longitudinal extension of LCA allowing for the examination of the conditional probabilities of an individual being in a particular latent class, given their latent class membership at the previous time point. Thus, in addition to latent class probabilities and latent class-specific item endorsement profiles as described with the LCA model, LTA models also include a transition probability. The transition probability is the probability of an individual belonging to a particular latent class at a time point t, given their class membership at a time $t-1$.

Denote $\boldsymbol{y}_k$ by a vector of item responses for person j on all items I across time points T. The probability of $\boldsymbol{y}_k$ can be expressed as:

$$\Pr(Y_j \geq \boldsymbol{y}_k) = \left[\sum_{g_{(t=1)}=1}^{G} \Pr(C_{j(t=1)} = g_{(t=1)}) \right] \times \left[\prod_{t=2}^{T} \sum_{g_{(t-1)}=1}^{G} \sum_{g_t=1}^{G} \Pr(C_{jt} = g_t \mid C_{j(t-1)} = g_{(t-1)}) \right] \times \left[\prod_{t=1}^{T} \sum_{g_t=1}^{G} \prod_{i=1}^{I} \Pr(Y_{jit} \geq y_k \mid C_{jt} = g_t) \right], \quad (3)$$

where C_{jt} is a discrete latent class variable denoting latent class membership for person j at a time point t and g_t is a specific individual-level discrete latent class at a time point t. This equation includes three probabilities. Ordered in the same way as found in equation 3, these are: the probability of latent class membership at a time point 1, the probability of transitioning to latent class *g* at time t given latent class membership time $t-1$ (for $t = 2,...,T$), and the probability of observing each item response at each time point conditional on latent class membership, respectively. More specifically, the last

probability in equation 3 is the probability of person j having Y_{jit} higher than or equal to category score y_k on item i at time t given their latent class membership at that time. Similar to Equation 1, this can be further written as:

$$\Pr(Y_{jit} \geq y_k \mid C_{jt} = g_t) = \frac{\exp(\delta_{ikg_t} - Y^*_{jit})}{1 + \exp(\delta_{ikg_t} - Y^*_{jit})}, \quad (3)$$

where δ_{ikg_t} is the item- and latent class-specific threshold parameters at a time point t. Specifically, δ_{i1g_t} is the threshold between response category 0 and 1 & 2 for item i in latent class g at time t, and δ_{i2g_t} is the threshold between response category 0 & 1 and 2 for item i in latent class g at time t, when there are 3 response categories. Y^*_{jit} is a continuous latent response variable for person j on item i at time t. In the current dissertation, item threshold parameters were held constant across time in order to keep the interpretations of latent class memberships consistent.

In summary, for the current application utilizing two time points ($T = 2$), Mplus produces three sections of results from LTA: 1) latent class proportions at the first time point; 2) transition probabilities for moving from latent class g at the first time point to latent class g at the second time point; and 3) item threshold parameters (δ_{ikg_t}). As outlined by Nylund (2007), the application of a LTA model to the data proceeded in three steps. First, LCA model parameters were estimated separately at each time point to determine the appropriate number of latent classes necessary to model the discrete latent variable. The suitable number of latent classes was determined based on the LMR-LRT and BIC, as previously introduced in the section describing the LCA model. Similar to

simulation studies with the LCA model, the BIC has been shown to be superior in simulation studies at identifying the population model in longitudinal applications of LCA (Nylund et al., 2007). Second, transitions were explored based on cross-sectional results of each time point. This was done to get an indication of the amount and type of movement in the data as well as to inspect for measurement invariance in terms of whether the interpretation of the latent classes remained the same over time. Finally, LTA model parameters were estimated by regressing latent classes at a time point 2 on those at a time point 1. During this step, the latent transition proportions were estimated. As with LCA models, estimation of LTA model parameters in Mplus is done utilizing the EM algorithm.

Similar to LCA models, and as is true for all mixture models, the estimation algorithm used with LTA may converge on local, rather than global solutions. This occurs when the algorithm converges on a maximum log-likelihood value that is locally optimal in the parameter space but is not necessarily the global maximum. The use of multiple starting values from random locations in the parameter space is a common method for checking that convergence occurred at a global maximum. By observing the same maximum likelihood value across multiple random starting values, there is more confidence that the global maximum was obtained. The current application used 200 random start values with the 20 best optimizations. A solution was chosen if the final log likelihood value was replicated multiple times.

Another common problem seen in mixture modeling is that of label switching. This occurs when, for example, Latent Class 1 (e.g., Clinical Severity Class) at time 1 is also generated at time 2 but is labeled as Latent Class 2 in the output. Label switching

such as this does not impact the parameter estimates, but does have implications for correct interpretation of results. To ensure that there is no label switching occurred that would interfere with proper interpretation of latent classes, item thresholds, as well as mixture proportions were compared across time.

Model 4: Longitudinal Graded Response Model

As a longitudinal extension of the GRM, the longitudinal GRM (Muraki & Carlson, 1995) allows for the investigation of change over time by analyzing repeated administrations of the same items. Thus, the probability of person j having Y_{jit} higher than or equal to category score y_k on item i at time a point t is defined as:

$$\Pr(Y_{jit} \geq y_k \mid \theta_{jt}) = \frac{\exp(-\delta_{ikt} + \lambda_{it}\theta_{jt})}{1 + \exp(-\delta_{ikt} + \lambda_{it}\theta_{jt})}, \quad (4)$$

where δ_{ikt} indicates the item thresholds between the response categories of item i at time t, λ_{it} is the factor loading (see footnote 1) for item i at a time point t, and θ_{jt} is the trait level (i.e., symptom severity) of person j at a time point t. Thus, with two time points in this dissertation, θ_{j1} is the trait level for person j at a time point 1 and θ_{j2} is the trait level for person j at a time point 2. Therefore, a person's change in trait level between a time point 1 and a time point 2 can be computed as $\theta_{j2} - \theta_{j1}$ where a resulting negative number reflects a decrease in trait level (Andersen, 1985). In terms of the specific case of the SFSS Externalizing subscale, this decrease signifies a reduction in externalizing symptom severity. In this example with two time points, trait levels $(\boldsymbol{\theta}_j = [\theta_{j1}, \theta_{j2}]')$ are assumed to have a multivariate normal distribution $(\mathbf{0}, \sum)$ where:

$$\Sigma = \begin{bmatrix} 1 & \sigma_{21} \\ \sigma_{12} & 1 \end{bmatrix}.$$

An important assumption underlying the longitudinal GRM is that of measurement invariance across time. If this assumption is violated, observed changes in scores over time are ambiguous and difficult to interpret given they may be a result of change in item parameters over time rather than 'true' trait change. As described in Millsap (2010) and Pastor and Beretvas (2006), separate steps were taken to investigate measurement invariance. First, separate GRMs were applied to the data at each time point to ensure the model provided an adequate fit to the data at a time point 1 and a time point 2. The second step assessed measurement invariance by investigating differential item functioning (DIF) by time with item parameters (δ_{ikt} and λ_{it}). DIF assesses potential item bias between time points by assessing whether, when controlling trait level, there is a secondary latent dimension that is leading to the between-group differences in item parameters. If this secondary dimension of time is needed to describe dependency in item responses, item bias is present and the assumption of measurement invariance is violated. Three methods will be utilized for inspection of DIF based on time: the Generalized Mantel-Haenszel statistic (GHM; Zwick, Donoghue, & Grima, 1993), a Logistic Regression (LR; Miller & Spray, 1993; Zumbo, 1999) model, and likelihood-ratio procedure with concurrent IRT calibration. Detailed descriptions of these three methods are provided in Chapter 3 and the results are provided in Chapter 4.

Prior to presenting the results and discussion from application of these four models to the empirical data, the next chapter (Chapter 3) presents a thorough

psychometric evaluation of the Caregiver SFSS Externalizing subscale. This psychometric testing utilized methods from CTT, IRT, EFA, and CFA, to validate the measure for caregivers of clinically referred youth. Although the SFSS has been previously validated in this population (see Bickman et al., 2007), these results were based on using 5-point Likert-type item response options. The current psychometric evaluation assessed the SFSS Externalizing subscale with the reduced 3-point item response options. Therefore, the purpose of these psychometric analyses is to ensure the psychometric qualities of the measure hold when reduced response options are used. Additionally, Chapter 3 provides a more detailed investigation of this subscale, including assessment of DIF based on youth gender and age. Results of testing for the unidimensionality of the SFSS Externalizing subscale are also presented.

CHAPTER 3

Validation of the Externalizing Subscale of the SFSS

This dissertation used data collected from the caregiver version of the SFSS Externalizing Subscale to compare categorical and dimensional latent models of externalizing symptoms. Before adopting any measure for use, it is important that it be validated for its intended purpose and that it demonstrate adequate psychometric properties, including reliability and validity. Establishing unidimensionality is also essential to the application of IRT models, as previously discussed in Chapter 2. Therefore, the goal of this chapter is to present the results of a rigorous psychometric analysis of the caregiver version of the SFSS Externalizing subscale with reduced 3-response options. This psychometric study utilized methods from CTT , IRT, EFA, and CFA to examine individual item properties, scale reliability, construct validity, and the unidimensionality of the SFSS Externalizing subscale in a large sample (N = 668) of caregivers for clinically-referred youth.

Method

Participants

The participants included in the psychometric evaluation were drawn from the same CFS™ evaluation study described in Chapter 2. During the two and a half year data collection period after CFS™ was implemented, youth (and their respective

caregivers and clinicians) who started treatment were entered into the system and began completing measures and contributing data. The first time point for these youth reflects the beginning of treatment. However, when CFS™ was implemented, youth who were currently receiving treatment also began contributing data within the system. Their first time point does not reflect the beginning of treatment, but rather some point within their treatment process. The psychometric evaluation included all caregivers who completed an SFSS Externalizing subscale at any time during CFS™ data collection. If caregivers completed more than one SFSS Externalizing subscale, scores for the first completed one were used. Data were received de-identified after a rigorous data processing protocol (see Bickman et al., 2007, 2010). The Institutional Review Board of Vanderbilt University granted approval of data collection.

Measures

Measures listed here were used within CFS™ and are part of the first edition of the PTPB (Bickman et al., 2007). For all measures, completion was defined as having 80% non-missing item responses. If more than 20% of item responses were missing, a total score was not computed and was instead reported as missing. If fewer than 20% of items were missing, mean imputation was used for missing item responses. For respondents who completed any given measure more than once, the first one was used.

SFSS Externalizing subscale: As previously described, the SFSS Externalizing subscale is composed of 16 Likert-type items rated 1 to 5 (*never, rarely, sometimes, often, very often*). The Externalizing score in this chapter is computed as the mean of responses to items about the frequency, in the last two weeks, the youth had engaged in

each of the externalizing symptom/behaviors. As previously mentioned, each item's response categories were collapsed into three categories (*never, rarely/sometimes, often/very often;* coded 0, 1, 2) and the total score computed accordingly.

Satisfaction with Life Scale (SWLS). The SWLS was developed by Diener, Emmons, Larson & Griffin (1985) and is the most popular scale for measuring life satisfaction (Diener, Suh, Lucas & Smith, 1999; Vassar, Ridge & Hill, 2008). The five items of the SWLS are: '*In most ways my life is close to my ideal*'; '*The conditions of my life are excellent*'; '*I am satisfied with my life*'; '*So far I have gotten the important things I want in my life*'; and '*If I could live my life over, I would change almost nothing*'. Respondents are asked to answer each item on a 7-point Likert scale (from 1= *strongly disagree* to 7 = *strongly agree*). Item responses are averaged to create a summary score from 1-7. Pavot & Diener (2008) reported an average item score of 4 as neutral, > 6.2 indicating '*extremely satisfied*' and < 2 as '*extremely dissatisfied'*. The SWLS has a reported Cronbach's alpha of .87, a test-retest correlation of .82, and a single factor solution has been replicated through factor analysis (Diener et al. 1985; Neto, 1993). The psychometric properties of the SWLS have also been established as adequate for use with caregivers of clinically referred youth (Athay, 2012).

Caregiver Strain Questionnaire-Short Form 10 (CGSQ-SF10). Composed of ten items from the original 21-item CGSQ (Brannan, Heflinger & Bickman, 1997), the CGSQ-SF10 assesses the extent that caregivers experience objective and subjective strain from caring for a child with mental health difficulties. It yields a total score and two subscale scores (objective and subjective strain). The CGSQ-SF10 displays excellent psychometric properties including a Cronbach's alpha of .90 (Bickman et al., 2007).

Treatment Outcomes and Expectations Scale (TOES). The TOES assesses youth and caregiver's expectations about the anticipated outcome of treatment. Composed of eight items, the TOES is completed by the youth and caregiver at the beginning of treatment. The TOES displays excellent psychometric properties including a Cronbach's alpha of .91 for the youth version and .85 for the caregiver version. See Dew-Reeves and Athay (2012) for more information about the TOES.

Service Satisfaction Scale (SSS). The SSS provides a general indicator of how well the caregiver perceives the mental health organization's service. It demonstrates adequate psychometric properties including a Cronabch's alpha of .85. For more information about the SSS, see Athay and Bickman (2012).

Background questionnaire. At treatment baseline, youth and caregivers completed background questionnaires. These forms collected background information such as gender, age, and ethnicity.

Analytic Approach

For the evaluation of the psychometric properties of the SFSS Externalizing Subscale, methods from CTT and IRT were used. Together, results from CTT and IRT methods provide important insights concerning psychometric qualities of individual items as well as the overall scale. However, each of these models has their own strengths and weaknesses. The strength of CTT includes its ease of use and wide familiarity among most readers. However, the resulting statistics are sample dependent and include arithmetic operations that require variables measured at an interval scale level. Unfortunately, interval level scaling is not empirically proven for rating scale items. IRT,

on the other hand, is able to provide more detailed item-level information that is less sample-dependent, while also being able to create linear interval-level scales (Embretson, 1996). This is accomplished by utilizing a model that estimates both item-level and person-level parameters on a logit scale. Thus, items and persons are ordered along the same latent trait continuum. Although several different item response models have been developed, the rating scale model (RSM) with polytomously scored items (Andrich, 1978) was use for current psychometric purposes. RSM analyses were conducted with ConQuest software (Wu, 2007).

Item properties. Within CTT and IRT, individual items can be described based on their difficulty (called "item severities" here) and discrimination. Generally, item severities indicates the rarity of endorsement, where one would expect only individuals with high externalizing severity to endorse a high severity item and individuals both with high and low severity to endorse a very low severity item. Item discrimination refers to the ability of an item to discriminate between respondents with high severity and those with low severity. Items without the ability to discriminate contribute little or no measure information.

Within the CTT framework, item difficulties were calculated using mean score responses. Items with extremely low or extremely high mean scores are items that too few or too many people endorse. When that occurs, items contribute little information to a scale. Item discrimination within CTT is often expressed statistically with the Pearson product-moment correlation coefficient (Pearson's *r*) between the item and total scores.

Within the IRT framework, item location parameters are estimated as well as their associated standard errors, and mean square fit statistics (MNSQ). In the current

application item locations are called item severities. Item severities show where an item is most precise in estimating a person's trait level. This can be depicted in a Wright map where items and persons are plotted on the same continuum. It is desirable for a measure to contain items located along the entire range of the latent trait. The MNSQ is an indicator how well an item fits the model. According to Wright and Linacre (1994), items with MNSQ between 0.6 and 1.4 contribute to the reliability of measurement and items outside that range do not. The RSM is a one-parameter model. Therefore, item discriminations are set equal across items and no individual item discriminations are computed.

Reliability. Reliability, or the degree to which a test is consistent in its measurement, is an important consideration for the use of measures. It is critical to have highly reliable measures in order to trust the resulting data. CTT and IRT methods provide slightly different ways to examine reliability but are rather similar in interpretation.

In CTT, the Cronbach alpha statistic is often used to report reliability (or internal consistency). This is the proportion of variance accounted for by the model and is based on item covariances. In psychology research, the general rule of thumb is for measures to have an alpha of at least 0.80 (Nunnally & Bernstein, 1994). Additionally, the standard error of the measurement (*SEM*) is used as an indication of reliability. It quantifies the amount of uncertainty there is around a score, where a smaller *SEM* indicates more precise, or consistent, measurement. A measure's *SEM* is the average of individuals' standard errors. Thus, one *SEM* is reported for a measure.

Reliability within IRT modeling can be reported with the separation reliability statistic (Wright & Masters, 1981 as reported in Wilson, 2005). This is the amount of total variance explained by the estimated person trait-level parameters. Although there are no steadfast rules or cut-off scores for determining acceptable separation reliability, values close to one are desirable. IRT also allows for the calculation of standard error of estimates for person trait scores. However, unlike CTT, these standard errors can vary across the latent trait continuum. Therefore, graphing of standard errors according to trait estimates allows for a visual inspection of where a measure is most precise in measurement across the continuum.

Construct Validity. Validity is another important feature of a measure. Construct validity refers to how well or to what degree a measure is actually measuring what it is purported to measure. Assessing construct validity can include investigating how well the measure corresponds with the theoretical ideas behind the trait, as well as how the scale correlates with variables known to be related or un-related to that trait. Additionally, as within the IRT framework, construct validity may include demonstrating items are unbiased for groups of individuals.

To assess construct validity under the CTT framework, both EFA and CFA were used. The SFSS Externalizing subscale was developed as a unidimensional scale measuring a single construct. Therefore, all item responses are combined to create one total scale score representing the respondent's level of externalizing severity. The interpretations made from this total score are valid as long as the assumption that the measure is unidimensional remains true. In the current sample, EFA was used to explore the factor structure, and CFA was used to test the unidimensional assumption by loading

all items on a single latent variable. In addition to providing evidence for construct validity, establishing unidimensionality is essential prior to IRT modeling. The SAS® procedure PROC CALIS was used for EFA and CFA analyses.

Under the CTT framework and consistent with the multitrait-multimethod matrix of examining construct validity (Campbell & Fisk, 1959), patterns of relationships between the SFSS Externalizing score and other variables were also inspected,. These other variables were chosen based on their theoretical and empirical relationships to externalizing symptom severity. Youth externalizing symptom severity has been shown to significantly relate to caregiver strain (Ekas & Whitman, 2010; Hastings, Daley, Burns, Beck & MacLean, 2006) and caregiver life satisfaction (Ekas & Whitman, 2010; Grosse, Flores, Ouyang, Robbins & Tilford, 2009), but no theoretical or empirical evidence demonstrates a relationship with caregiver service satisfaction or treatment expectations. Therefore, it is expected that SFSS Externalizing scores will significantly correlate with CGSQ-SF and SWLS total scores and will not significantly correlate with SSS and TOES total scores.

Construct validity within the IRT framework was assessed in terms of DIF, the presence of which can directly influence an instrument's validity. It is important to note that DIF is distinct from the differential impact of items seen within subgroups. For example, it may be that males typically score higher than females on a particular measure or item. This difference does not influence validity. Within IRT, on the other hand, validity is affected if males and females with the same trait level respond differently to items. This would indicate DIF based on gender, meaning items are biased. Youth gender and age are two standard grouping variables investigated for differences in

externalizing symptom severity. Therefore, DIF was investigated for the SFSS Externalizing items based on youth gender and youth age.

Methods for DIF analysis require the division of participants into two groups: the reference group and the focal group. The focal group is the group believed to be disadvantaged by an item, and the reference group is the standard that the focal group is compared to. Thus, youth in these two groups were matched based on their level of externalizing symptom severity (either their observed score or latent score) and group differences were then analyzed using one of many statistical procedures, three of which will be demonstrated in the current chapter. These procedures will also be used to test for invariance over time for longitudinal IRT, as previously mentioned in Chapter 2. Descriptions of the grouping variables used in DIF analyses for the psychometric analyses are found in Table 4.

Table 4. Grouping Variables used as Focal Group in DIF Analyses

Grouping Variable (Focal Group)	Description
Female Gender	Youth are Female (1), male (0)
Younger youth	Youth are aged 11-12 (1), else (0)
Older youth	Youth are aged 16-18 (1), else (0)

Millsap and Everson (1993) categorized DIF procedures into two categories: 1) observed conditional invariance (OCI) procedures, and 2) unobserved conditional invariance (UCI) procedures. OCI procedures match individuals in the reference and focal group based on observed total scores whereas UCI procedures match individuals

based on latent, or unobserved, trait scores. Two OCI and one UCI method will be utilized in the current study: the Generalized Mantel-Haenszel statistic (GHM; Zwick et al., 1993) and Logistic Regression (LR; Zumbo, 1999) model for the OCI procedures and concurrent IRT calibration for the UCI approach. OCI methods were conducted with SAS® version 9.2 software and the UCI method utilized *ACER ConQuest version 2.0* (Wu, 2007).

The GMH statistic tests the conditional independence for a grouping variable and an item by assessing between-group differences in the frequencies of the item scores when the total score is controlled (Zwick et al., 1993). Nominal numbers are assigned to the response categories and item response vectors for individuals in the reference and focal group are compared after being matched on their observed (total) score. The LR procedure tests the difference in deviance statistics for three related models: the full model and two reduced models (Zumbo, 1999). The full model predicts the probability of an item response with the total score, the grouping variable, and the interaction between the total score and the grouping variable. The first reduced model predicts the probability of an item response with the total score and grouping variable, and the second reduced model predicts the probability of an item response only with the total score. Comparison of deviance statistics for the full and first reduced model allows for inspection of potential non-uniform DIF, whereas comparison of deviance statistics for the two reduced models allows for inspection of potential uniform DIF. Non-uniform DIF exists when there is an interaction between the total score and group membership such that the between-group difference in the probability of an item response is not the same across all levels of externalizing symptom severity. Uniform DIF exists when there

is no interaction between the total score and group membership. That is, the probability of an item response is greater for one group over the other in a uniform fashion over all levels of externalizing symptom severity.

The concurrent IRT calibration method is a UCI approach for the investigation of DIF. For this method, Rasch model item severities were estimated separately for individuals within a category of a grouping variable. By plotting the resulting item severities of the reference group against the focal group, the comparability of item severities between groups can be visually inspected. For items that have similar severities across groups (i.e., no DIF is present), item plots will fall on a 45-degree line. Items with potential DIF will fall away from the 45-degree reference line, indicating potential item bias.

Results

Demographics of the Current Sample

The total sample included 668 caregivers and their respective clinically referred youth. Caregivers ranged in age from 23 to 81 years (M = 44.7; SD = 10.54) and youth ranged in age from 11-18 years (M = 14.7; SD = 1.84). Slightly more than half of the youth were male (54%) and approximately 86% of caregivers were female. The majority of caregivers reported being the youth's primary caregiver (96%) who live with the youth full time (97%). Ethnic breakdown of caregivers and youth who reported their racial background can be found in Table 5.

Table 5. Racial Background of Caregivers and Youth

	Caregivers Total *N* = 375		Youth Total *N* = 402	
Racial Background	*N*	Percent	*N*	Percent
African American	111	29.6	106	26.4
American Indian or Alaskan Native	7	1.9	5	1.2
Asian	4	1.1	2	0.5
Caucasian or white	225	60.0	207	51.6
Native Hawaiian/Pacific Islander	2	0.5	1	0.2
More than one race	7	1.9	47	11.7
Other	19	5.1	34	8.5

Note: Missing: Caregivers *N* = 293; Youth *N* = 266

Item Properties

Descriptives from on CTT analyses of SFSS Externalizing items and total score can be found in table 6. The distribution of the total Externalizing score in the current sample had a mean of 1.00 and a standard deviation of 0.49. Pearson *r*'s ranged from .52 to .75. Item 10 ('*hangs out with peers who get in trouble*') had the lowest discrimination (*r* = .52), indicating it may not be able to distinguish between those with high and low externalizing severity compared to the other items. Kurtosis and skewness values indicated neither the items nor total score were excessively leptokurtic or skewed (Harlow, 2005, p. 34).

Table 6. Descriptive Statistics SFSS Externalizing Items and Total Score (N = 668)

		CTT					RSM		
Item No.	Description	Mean	*SD*	skewness	kurtosis	Disc.	Location	*SE*	MNSQ
1	Throw things	0.70	0.69	0.48	– 0.83	0.66	1.15	0.05	1.10
2	Get in trouble	1.02	0.67	– 0.02	– 0.79	0.74	– 0.08	0.05	0.82
3	Disobey adults	1.21	0.66	– 0.26	– 0.77	0.78	– 0.82	0.05	0.73
4	Interrupt others	1.15	0.67	– 0.18	– 0.79	0.72	– 0.58	0.05	0.87
5	Lie to get things	0.99	0.73	0.02	– 1.12	0.70	0.04	0.05	1.09
6	Can’t control temper	1.18	0.70	– 0.27	– 0.94	0.74	– 0.69	0.05	0.93
7	Hard to get along	1.06	0.67	– 0.07	– 0.75	0.72	– 0.19	0.05	0.88
8	Threaten/bully	0.70	0.70	0.50	– 0.89	0.68	1.13	0.05	1.10
9	Hard to wait turn	0.84	0.70	0.23	– 0.97	0.69	0.60	0.05	1.05
10	Troubled peers	0.73	0.73	0.47	– 1.02	0.58	1.00	0.05	1.43
11	Can’t pay attention	1.16	0.68	– 0.21	– 0.83	0.67	– 0.67	0.05	1.01
12	Get into fights	1.00	0.70	0.01	– 0.95	0.76	0.02	0.05	0.86
13	Lose things	1.00	0.70	0.00	– 0.93	0.64	0.04	0.05	1.14
14	Cant’ sit still	1.05	0.71	– 0.08	– 0.99	0.64	– 0.19	0.05	1.17
15	Annoy others	1.06	0.71	– 0.08	– 1.00	0.71	– 0.23	0.05	0.98
16	Argue	1.15	0.73	– 0.23	– 1.10	0.78	– 0.53	*	0.88
Total Mean Scale Score		1.00	0.49	– 0.65	– 0.13	– 0.65	--	--	--

SD = standard deviation; Disc. = Discrimination as Pearson *r;* Location = item severity; *SE* = standard error; MNSQ = mean square statistic
*parameter estimate constrained

Estimated location parameters, their associated standard errors, and the weight fit statistics from RSM analyses are also found in Table 6. Item locations ranged from −0.69 to 1.15 on a logit scale. Endorsement of item 1 ('*throws things when mad*') indicated the highest level of externalizing symptom severity and item 6 ('*difficulty controlling temper*') indicated the lowest. This is depicted in the Wright map (see Figure 6). The Wright map places all items (and persons – not included in Figure 3) on the same latent continuum. As can be seen, all items are relatively clumped together at the center of the continuum, with some overlap. This indicates that the SFSS Externalizing subscale is most precise for measurement at the center of the continuum. This is discussed later in terms of reliability.

Item thresholds were at −1.53 and 1.53 logit units. This means that a person with a trait score of −1.53 logit units is just as likely to endorse response category 0 (*never*) as they would endorse response category 1 (*rarely/sometimes*). Given these thresholds are in the expected order and are spaced apart sufficiently, it appears that these three response categories are distinguishable from one another.

According to MNSQ statistics, most items were within the range for acceptable model fit (i.e., between 0.6 and 1.4; Wright & Linacre, 1994). However, item 10 ('*hang out with peers who get in trouble*') was slightly elevated. This means some caregivers endorsed this item in unexpected ways, or that this item measures unmodeled variance (i.e., noise).

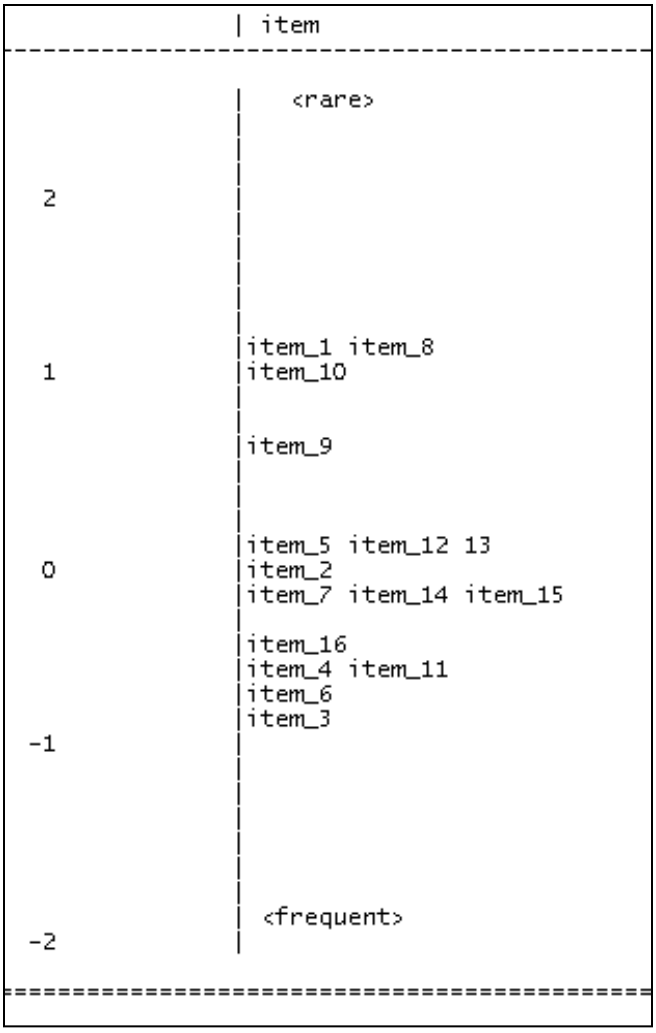


Figure 6. Wright map of items (generated by *ConQuest*; Wu, 2007).

Reliability

Reliability within CTT was reported with Cronbach's alpha. The SFSS Externalizing subscale had an alpha of .93, demonstrating adequate internal consistency. Reliability within IRT modeling was quantified by separation reliability. As previously noted, there are no steadfast rules or cut-off scores for acceptable separation reliability. However, the value for the Externalizing subscale was .99, which is near the highest possible score of 1.0. It is safe to conclude that the SFSS Externalizing subscale has

adequate reliability within the IRT framework. Graphical results of the calculated person-trait estimates and their associated standard errors are found in Figure 7. As can be seen, the SFSS Externalizing subscale is most accurate for measuring of externalizing symptom severity at the center of the latent trait continuum (approximately between –1.0 and 1.0) and less accurate farther away from the center.

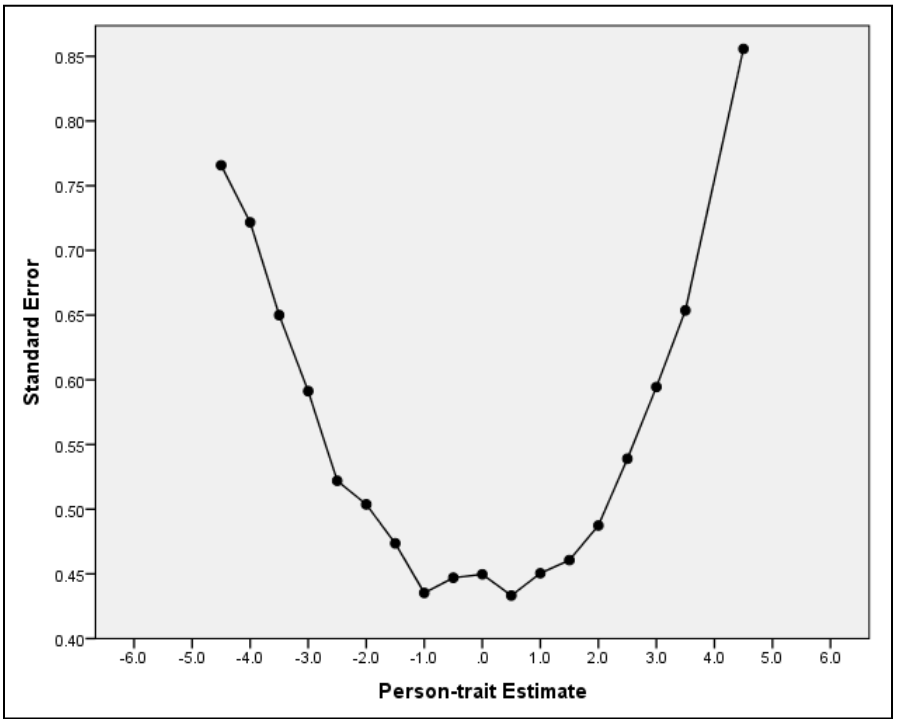


Figure 7. Calculated person-trait estimates and their associated standard errors

Construct Validity

Responses to the SFSS Externalizing subscale were analyzed with EFA using principal components. The eigenvalue-one criterion (Kaiser, 1960) was used to determine the number of factors to retain. This criterion (also known as the Kaiser criterion) retains any eigenvalues greater than one. According to this criterion, two

factors were retained in the current analysis, with eigenvalues of 8.02 and 1.19 respectively. These results are depicted in Figure 8 as a scree plot.

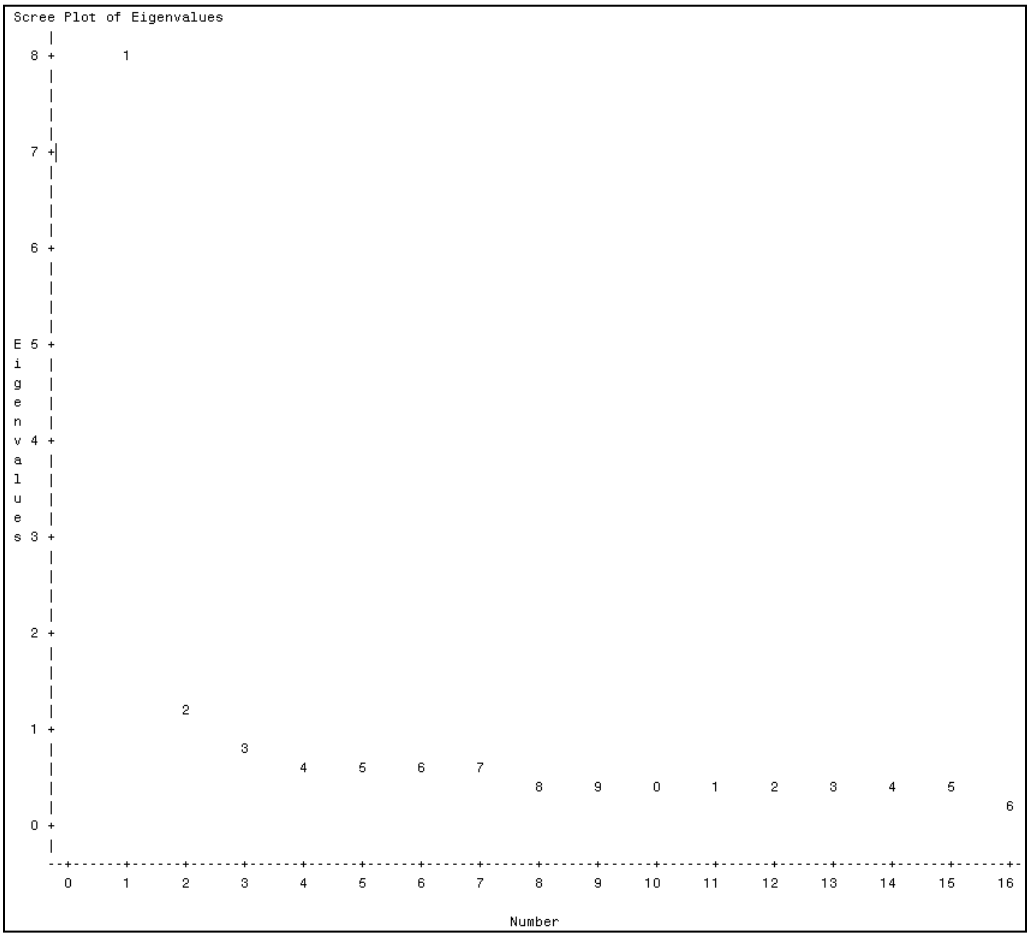


Figure 8. Scree plot of eigenvalues for the SFSS Externalizing Subscale

To assess model fit within CFA, three popular fit statistics were used: Bentler's Comparative Fit Index (CFI; Bentler, 1990), Joreskog's Goodness of Fit Index (GFI; Joreskog, 1988**),** and the Standardized Root Mean Square Residual (SRMR; Steiger, 2000). Results indicated that the proposed unidimensional model fit the data slightly less than commonly agreed upon standards. According to Browne and Cudeck (1993), values

greater than 0.90 indicate good fit between a model and the data for the CFI and GFI. For the SRMR, a value of 0.05 indicates close fit, 0.08 fair fit, and 0.10 marginal fit (Hu & Bentler, 1999). The current results found GFI = 0.87, Bentler's comparative fit index = 0.90, and RMSEA = 0.09 for the SFSS Externalizing subscale.

As hypothesized, the SFSS Externalizing score significantly correlated with caregiver strain (CGSQ-SF10; $r = .55$, $p < .001$) and caregiver life satisfaction (SWLS; $r = -.27$, $p < .001$), with higher externalizing symptom severity related to higher caregiver strain and lower life satisfaction. Additionally, the SFSS Externalizing score was not significantly correlated with outcome expectations (TOES; $r = -.09$, $p = .16$) or caregiver service satisfaction (SSS; $r = -.09$, $p = .08$). Together, these results suggest the construct of externalizing symptom severity (as rated by the SFSS) relates to these other constructs in similar and expected ways according to theory and previous research.

Results of OCI DIF procedures are found in table 7. Based on youth gender, the GMH statistic indicated potential DIF for item 7 ($MH\chi^2 = 18.27$, $p < .05$), item 12 ($MH\chi^2 = 6.90$, $p < .05$), and item 14 ($MH\chi^2 = 16.61$, $p < .05$). For the youngest youth (aged 11-12), potential DIF was also indicated for item 5 ($MH\chi^2 = 12.25$, $p < .025$), item 9 ($MH\chi^2 = 17.43$, $p < .025$), item 10 ($MH\chi^2 = 12.40$, $p < .025$) and item 14 ($MH\chi^2 = 15.60$, $p < .025$). Additionally, the GMH statistic indicated potential DIF for item 6 ($MH\chi^2 = 8.32$, $p < .025$), item 9 ($MH\chi^2 = 8.05$, $p < .025$), and item10 ($MH\chi^2 = 22.10$, $p < .025$) for older youth. To investigate the strength of DIF, the probabilities for persons in the focal group indicating a different response category for an item compared to a person in the reference group were calculated when persons were matched across category by SFSS Externalizing score. These probabilities are found in Table 8. Only one item had a

probability greater than .75. This was item 5 (‘*lies to get things*’) for younger clients (aged 11–12). Thus, when matched for overall Externalizing SFSS score, the probability that a caregiver for a younger youth chose a different response category for this item compared with a caregiver of an older youth was .85. This may mean that item 5 is located on a different place on the latent continuum, depending on the age of the youth. The probabilities for all other items indicated by the GMH procedure for potential DIF were less than .75, with the majority of them falling at or below chance levels.

Results of LR analyses indicated potential uniform DIF based on gender for item 7 ($\Delta G^2 = 13.54$, $p < .025$), item 12 ($\Delta G^2 = 6.67$, $p < .025$), and item 14 ($\Delta G^2 = 13.69$, $p < .025$). Potential uniform DIF for younger youth was also indicated for item 4 ($\Delta G^2 = 5.18$, $p < .025$), item 5 ($\Delta G^2 = 15.28$, $p < .025$), item 9 ($\Delta G^2 = 15.31$, $p < .025$), item 10 ($\Delta G^2 = 11.79$, $p < .025$), item 12 ($\Delta G^2 = 5.04$, $p < .025$), and item 14 ($\Delta G^2 = 13.81$, $p < .025$). Additionally, seven items were indicated as having potential DIF for older caregivers: item 5 ($\Delta G^2 = 5.37$, $p < .025$), item 6 ($\Delta G^2 = 8.69$, $p < .025$), item 7 ($\Delta G^2 = 7.72$, $p < .025$), item 9 ($\Delta G^2 = 12.25$, $p < .025$), item 10 ($\Delta G^2 = 27.81$, $p < .025$), item 11 ($\Delta G^2 = 7.48$, $p < .025$), and item 14 ($\Delta G^2 = 12.07$, $p < .025$). Calculating effects sizes allows for the interpretation of magnitude for the potential DIF items. Zumbo (1999) proposed calculating two measures of magnitude by looking at the difference between the two reduced models in terms of their generalizing coefficients of determination and the coefficients of determination rescaled by their maximum values. These values are in Table 9, all of which fall well below the proposed cutoff value of 0.13 (Zumbo, 1999). Although the LR test appears sensitive to differences in item functioning between groups, the magnitude of these differences were negligible.

Table 7. Results of OCI DIF Analyses for SFSS Externalizing Items

Item and Description	DIF by Gender			DIF by Youngest			DIF by Oldest		
	GMH	non LR	uni LR	GMH	non LR	uni LR	GMH	non LR	uni LR
1: Throw things when mad	0.87	0.18	0.59	3.47	3.39	0.90	0.29	2.94	0.01
2: Get in trouble	0.33	0.00	0.01	6.74	2.06	4.54	1.50	0.21	1.01
3: Disobey adults	0.12	0.97	0.05	1.77	0.99	0.41	0.98	0.08	0.80
4: Interrupt others	0.78	0.90	0.50	7.02	0.18	**5.18**[B]	2.57	0.03	4.47
5: Lie to get things	4.64	0.03	3.78	**12.25**	0.48	**15.28**[B]	5.66	0.58	**5.37**[B]
6: Hard to control temper	2.49	2.28	0.01	1.81	0.14	0.89	**8.32**[A]	0.00	**8.69**[B]
7: Not get along fam/friend	**18.27**[A]	1.82	**13.54**[B]	5.02	0.17	4.02	7.26	0.02	**7.72**[B]
8: Threaten/bully others	3.36	0.00	2.96	1.09	0.54	0.56	0.01	2.20	0.01
9: Can't wait turn	3.94	0.13	2.69	**17.43**[A]	0.31	**15.31**[B]	**8.05**[A]	0.01	**12.24**[B]
10: Hang with troubled peers	2.68	3.58	1.60	**12.40**[A]	1.04	**11.79**[B]	**22.10**[A]	0.03	**27.80**[B]
11: Can't pay attention	3.72	1.78	1.54	0.33	0.15	0.14	6.72	3.11	**7.48** [B]
12: Gets into fights	**6.90**[A]	0.03	**6.67**[B]	7.13	0.34	**5.04**[B]	0.61	0.02	0.04
13: Loses things	5.67	1.45	0.01	3.13	1.66	0.88	7.10	2.23	3.73
14: Can't sit still	**16.61**[A]	1.09	**13.69**[B]	**15.60**[A]	0.01	**13.81**[B]	6.29	0.01	**12.07**[B]
15: Annoy others on purpose	2.09	0.77	1.52	2.74	2.98	0.31	2.86	0.02	3.50
16: Argues with adults	0.83	0.02	0.53	1.94	0.85	0.23	0.19	0.12	0.97

GMH = Generalized Mantel-Haenszel statistic; bolded if significant at $p < .050$ for gender, $p < .025$ for Youngest and Oldest

LR = Logistic Regression: (non) = Non-uniform DIF, (uni) = Uniform DIF; bolded if Chi square of difference in deviance statistics, df = 1, sig at $p < .05$ for gender, $p < .025$ for Youngest and Oldest.

[A] Probability of 0.75 or less of focal group responding different; [B] Negligible effect size = 0.05

Table 8. GMH Probabilities for Different Item Responses by Category

Gender		Younger		Older	
Item	Probability	Item	Probability	Item	Probability
7	.26	5	.84	6	.40
12	.40	9	.26	9	.58
14	.74	10	.71	10	.31
		14	.26		

Table 9. Effect Sizes for DIF items by LR Method

Gender			Younger			Older		
Item	Ch. in R^2	Ch. in Max R^2	Item	Ch. in R^2	Ch. in Max R^2	Item	Ch. in R^2	Ch. in Max R^2
7	0.012	0.014	4	0.004	0.005	5	0.004	0.005
12	0.005	0.006	5	0.012	0.014	6	0.006	0.007
14	0.015	0.017	9	0.012	0.014	7	0.006	0.007
			10	0.012	0.014	9	0.010	0.011
			12	0.003	0.004	10	0.029	0.033
			14	0.013	0.015	11	0.006	0.007
						14	0.011	0.013

Ch. = Change; R^2 = coefficient of determination; Max = rescaled to maximum value
Note: Values over 0.13 indicate significant DIF (Zumbo, 1999)

Use of the UCI scatterplot approach to investigate DIF is useful for obtaining a graphical representation of how items function across groups, when controlling for latent trait scores. The scatterplots of SFSS Externalizing items by each grouping variable are found in Figures 9–11. The solid line found on each graph is the 45-degree reference line where items without DIF are expected to fall. An item falling significantly above the reference line indicates that the item has a higher item severity (i.e., item location) for the reference group compared to the focal group. An item falling significantly below the line corresponds to a higher severity for the focal group as compared to the reference group. A difficulty in this approach is the lack of ability to determine how far from the reference line an item must fall to indicate DIF. To facilitate this however, confidence intervals of

approximately two standard errors for each item parameter are drawn for each item. Items with confidence intervals that fail to cross the reference line would be selected for further DIF investigation.

The scatterplot based on youth gender (Figure 9) suggests that items 5, 7, 8, 9, 10, 14, and 15 may have potential DIF. According to these results, endorsement of items 8, 9, 14, and 15 indicated a higher level of externalizing severity for girls compared to boys and endorsement of items 5, 7, and 10 indicated a lower level of externalizing severity for girls. Investigating DIF based on the youngest youth (Figure 10) found that endorsement of items 5, 7, 10, and 12 indicated higher level of externalizing severity for younger youth (aged 11–12) compared with youth older than 12, and endorsement of items 4, 9, and 14 indicated a lower level of externalizing severity for younger youth compared with those over 12. Finally, according to scatterplot results based on older youth (aged 16 – 18; Figure 11), endorsement of items 9, 11, 14, 15, and 16 indicated a higher level of externalizing severity for older youth compared to youth under age 16. Additionally, endorsement of items 5, 6, 7, and 10 indicated a lower level of externalizing severity for older youth compared with those under 16.

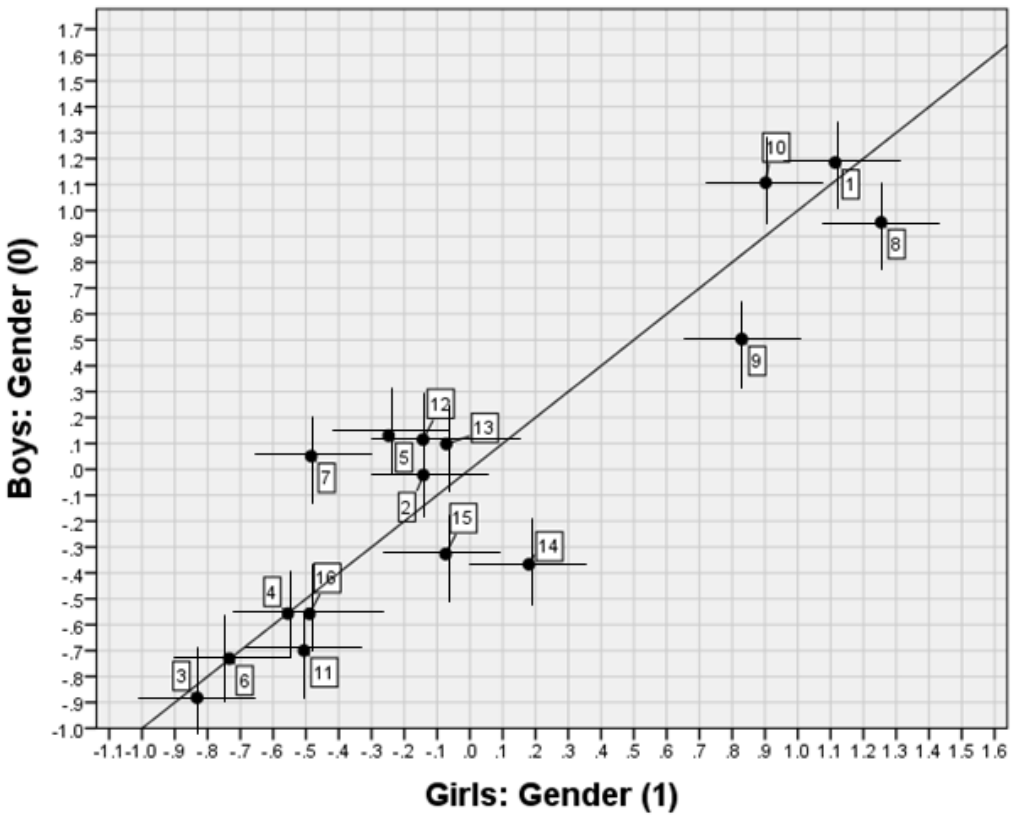


Figure 9. UCI scatterplot results: Comparison of item severities by gender.
Note: Confidence intervals drawn around +/- 2*SE* for item severity parameter estimates.

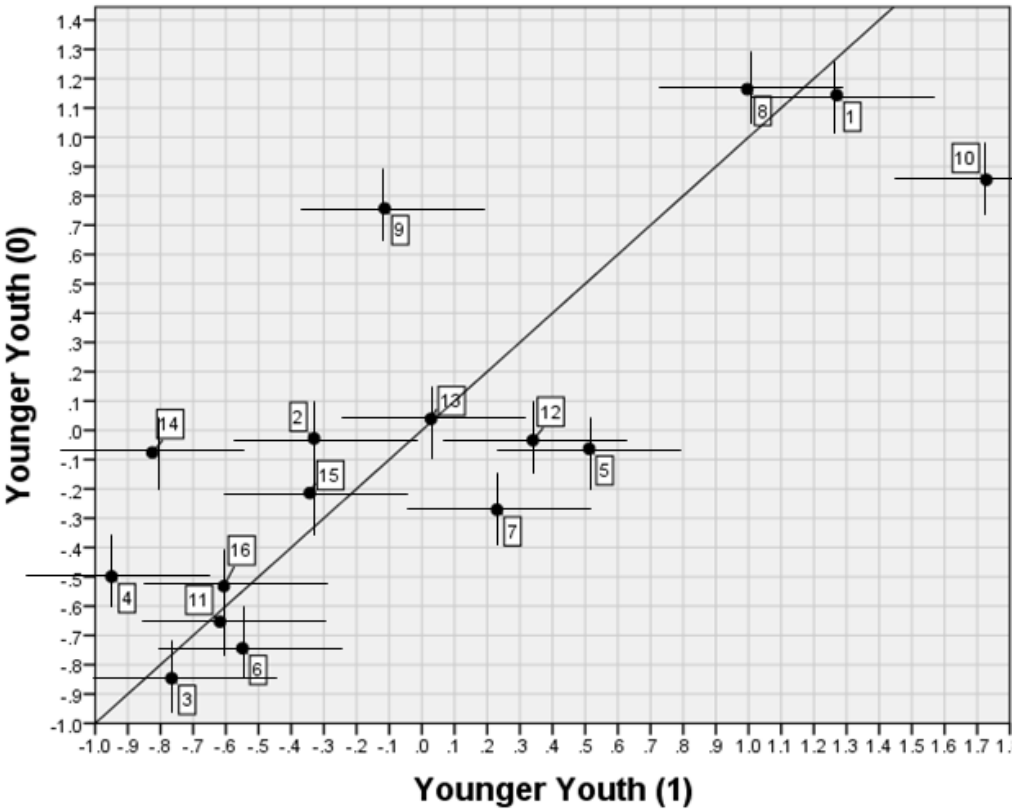


Figure 10. UCI scatterplot results: Comparison of item severities by Younger category.
Note: Confidence intervals drawn around +/- 2*SE* for item severity parameter estimates.

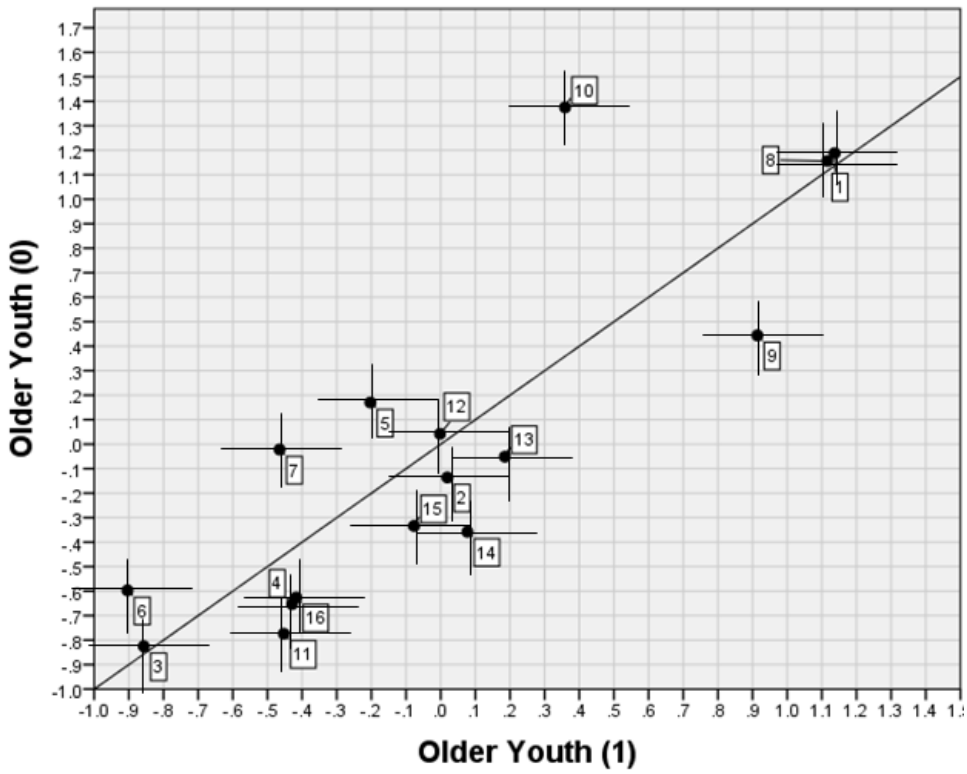


Figure 11. UCI scatterplot results: Comparison of item severities by older category.
Note: Confidence intervals drawn around +/- 2*SE* for item severity parameter estimates.

Discussion

In this chapter, the psychometric properties of the SFSS Externalizing subscale were evaluated in a large sample of caregivers for clinically referred youth aged 11-18. Overall, the results suggest the psychometric properties of the SFSS Externalizing subscale are satisfactory for this population. Scale scores and individual items were approximately normally distributed without significant kurtosis or skewness. Application of the Rasch measurement model indicated that the items fit the Rasch rating scale reasonably well although item 10 (‘*hangs out with peers who get in trouble*’) was slightly outside the desired range according to the MNSQ statistic. Future work is needed to investigate whether this item is problematic to the measure as a whole and should be

removed. However, given the otherwise adequate item properties and the desire to keep SFSS Externalizing subscale items parallel across respondent forms (i.e., versions for youth, caregivers, and clinicians; see Bickman et al., 2007), this item was retained.

In terms of reliability, the SFSS Externalizing subscale had high internal consistency, adequate item-total correlations, and high separation reliability. However, one potential weakness of this subscale was illuminated based on the item severities from the RSM analyses. The items of this subscale were located somewhat near each other in the middle of the latent continuum, with some overlap. This indicates that the SFSS Externalizing subscale is more precise at measuring externalizing symptom severity in the middle of the continuum and less precise at the tails. For precise measurement of the latent variable along the entire dimension, it is desirable to have the items spread out evenly over the continuum. However, the clumping of items on one portion of the continuum is common in psychological measurement and presents some unique challenges for clinical measurement within the IRT framework (Reise & Waller, 2009).

CFA fit indices (GFI, RMSEA, Bentler's CFI) indicated a less than ideal fit of the data to a one-factor model. Additionally, EFA indicated a two-factor solution, the first of which explained most of the variance. The less than ideal CFA and EFA fit is likely due to additional systematic variation related to the different diagnostic categories covered by the SFSS Externalizing subscale (ADHD, ODD, and CD). However, including these symptom categories as additional factors did not provide a better fit. Consistent with expectation, the SFSS Externalizing subscale significantly related to caregiver strain and caregiver life satisfaction and was unrelated to service satisfaction or treatment outcome

expectations. Taken together, these results provide some indirect evidence and support for the proposed unidimensional structure, especially given the high internal reliability.

The question of measurement invariance is important to address given its potential threat to construct validity. In order for a measure to be unbiased, it must measure externalizing symptom severity in the same way for different groups of youth. Therefore, to address the question of measurement invariance, DIF analyses were used. Across all categorical variables used in this study (gender, younger, older), the GMH analyses yielded a total of 10 items with potential DIF, the LR technique yielded a total of 16 items with potential DIF, and inspection of the UCI scatterplots identified 23 items with potential DIF. However, effect size estimation for GMH and LR approaches indicated small or insignificant DIF effects for these items. Furthermore, the UCI scatterplot approach lacks guidelines for determining effect sizes. Still, items identified across all three techniques as having potential DIF may warrant further investigation. These were items 7 and 14 based on gender, items 5, 9, 10, and 14 based on youngest youth and items 6, 9, and 10 based on older youth. However, given the small or negligible effect sizes calculated based on the OCI techniques, it is reasonable to conclude that there is measurement invariance across youth gender and age at this time. Further work is needed to confirm this.

As a whole, the results presented in this chapter provide evidence that the SFSS Externalizing subscale is both reliable and valid for assessing caregiver-rated youth externalizing symptom severity. However, further validation research is needed given that validation is a never-ending and circular process (Hubley & Zumbo, 1996).

Additionally, further analyses are needed to evaluate the measure's predictive validity and sensitivity to change in this population.

The purpose of this chapter was to establish the 3-response category SFSS Externalizing subscale as a psychometrically sound measure for use in this population. Additionally, this chapter presented some evidence supporting the proposed unidimensional structure of the measure, a preliminary step necessary for IRT modeling. The next chapter (Chapter 4), presents the results of fitting the four specified models of the empirical application, as described in Chapters 2 (i.e., LCA model, GRM, LTA model, Longitudinal GRM) to the data also described in Chapter 2. The final chapter, Chapter 5, presents the discussion.

CHAPTER 4

Results

The majority of this chapter is dedicated to presenting the results of applying the statistical models described in Chapter 2 to clinical data from the SFSS Externalizing Subscale completed by caregivers of youth receiving mental health treatment. These results apply directly to the second aim of this dissertation, which is to compare model results from the application of statistical models assuming different latent variable structures (i.e., categorical and dimensional). This type of comparison has not previously been done. The last part of this chapter is dedicated to presenting the results from the informal clinical survey described in Chapter 2. This, together with results from application of the statistical models, are used for the third aim of this dissertation: to demonstrate how the concept of cognitive fit proposes that certain presentations of clinical information support more effective and efficient decision-making depending on the nature of the specific decision being made.

LCA

Model Selection

To select the number of latent classes to use in the final model, LCA models with an increasing number of classes were fit to the data. Results are reported in Table 10. According to the BIC and LMR-LRT fit indices, the model with three latent classes

provided the best fit. Therefore, for the externalizing subscale of the SFSS, the model with three latent classes was selected as the final model.

Table 10. Fit Index Results for Model Selection (Time 1)

Number of Latent Classes	Number of Parameters	BIC[1]	LMR-LRT[2]
1	32	6587.49	
2	65	5782.38	975.04, $p < .001$
3	98	**5686.84**	**269.51, $p = 0.04$**
4	131	5701.74	157.99, $p = 0.08$

Note: BIC = Bayesian information criterion; LMR-LRT = Lo Mendell Rubin Likelihood Test
[1]Lowest number indicates best fit
[2]Significance indicates the model fits significantly better than a model with one fewer class

Item Parameter Estimates

The threshold estimates ($\hat{\delta}_{ikg}$) obtained from the 3-class LCA solution are found in Table 11. Each item has two estimated thresholds within each latent class. When an estimate approached an extreme, Mplus set it at −15 for the first threshold and 15 for the second. This means that in certain classes, the probability of item endorsement with a certain response category was zero or one, for −15 and 15, respectively. For example, the first threshold for item 1 in Class 1 was set at −15, which means that youth in this class have a probability of zero of endorsing this item below the first threshold (i.e., response category 1).

Table 11. Item Parameter Estimates from LCA Time 1

	Threshold 1						Threshold 2					
	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 1		Latent Class 2		Latent Class 3	
Item	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
1	−15.00	-	−1.37	0.28	1.64	0.35	−0.11	0.37	2.69	0.42	4.23	1.01
2	−15.00	-	−3.34	0.59	0.02	0.26	−1.79	0.53	1.24	0.27	4.20	1.01
3	−15.00	-	−15.00	-	−1.03	0.28	−2.39	0.62	−0.03	0.23	2.79	0.54
4	−15.00	-	−3.70	0.72	−0.79	0.27	−2.03	0.54	0.33	0.23	3.20	0.71
5	−15.00	-	−2.37	0.37	0.52	0.27	−1.09	0.44	0.37	0.22	2.53	0.48
6	−15.00	-	−15.00	-	−0.56	0.26	−2.90	0.92	0.13	0.22	2.54	0.48
7	−15.00	-	−3.43	0.59	−0.58	0.26	−1.87	0.62	0.93	0.24	2.83	0.53
8	−2.91	0.79	−1.08	0.26	1.29	0.30	−0.01	0.36	2.05	0.33	4.23	1.01
9	−15.00	-	−1.14	0.25	0.36	0.26	−0.70	0.41	2.02	0.34	3.79	0.95
10	−1.57	0.46	−0.74	0.24	0.61	0.26	0.27	0.36	1.25	0.26	4.22	1.01
11	−15.00	-	−2.54	0.40	−0.62	0.26	−1.72	0.55	0.35	0.22	2.44	0.47
12	−15.00	-	−2.92	0.56	0.13	0.25	−2.32	0.81	0.92	0.24	2.77	0.52
13	−2.42	0.71	−2.58	0.46	0.07	0.25	0.07	0.36	0.78	0.23	2.44	0.47
14	−3.51	1.02	−2.14	0.35	−0.37	0.26	−0.85	0.40	0.94	0.24	2.69	0.53
15	−15.00	-	−2.08	0.34	−0.52	0.26	−1.71	0.54	1.04	0.25	3.51	0.72
16	−15.00	-	−4.53	1.01	−0.29	0.26	−15.00	-	0.08	0.23	15.00	-

dash (-) = not calculated because estimate was fixed

Class Assignment

Assignment of participants to a latent class based on their largest posterior probability resulted in the proportion of class assignment depicted in Table 12. Class 2 was the largest with 98 youth (48%) and Class 1 was the smallest with 37 youth (18%). Class 3 had 69 youth (34%). Although the latent class assignment was based on the largest posterior probability for each youth, the average probability for these assignments was .96, .96, and .97 in Class 1, 2, and 3 respectively. The relative entropy, or the amount of classification certainty, was 0.92. This indicates a high amount of certainty in latent class assignment.

Table 12. Final Class Counts and Proportions (J = 204)

Latent Class	Frequency	Proportion (%)
1	37	18
2	98	48
3	69	34

Class Descriptions

The probability of endorsement of each SFSS externalizing item within each latent class was inspected in order to describe the classes according to externalizing symptom severity, as well as to qualitatively label the classes. These probabilities were calculated for each response category of every item and are specific to each latent class. Results are depicted in Figure 12; each panel represents a response category. For example, the bottom panel represents the probability within each latent class of endorsing each item at the highest response category, indicating the highest frequency/severity of the symptom. For all items in the bottom panel, the largest probability of endorsement at

this level was found for Class 1 and the smallest probability of endorsement was found for Class 3. The reverse was true in the top panel where, across items, Class 1 had the smallest probability of endorsement at the lowest response category, indicating no or low frequency/severity of the symptom, and Class 3 had the largest. For Class 2, the largest probability of item endorsement occurred mostly at the middle response category.

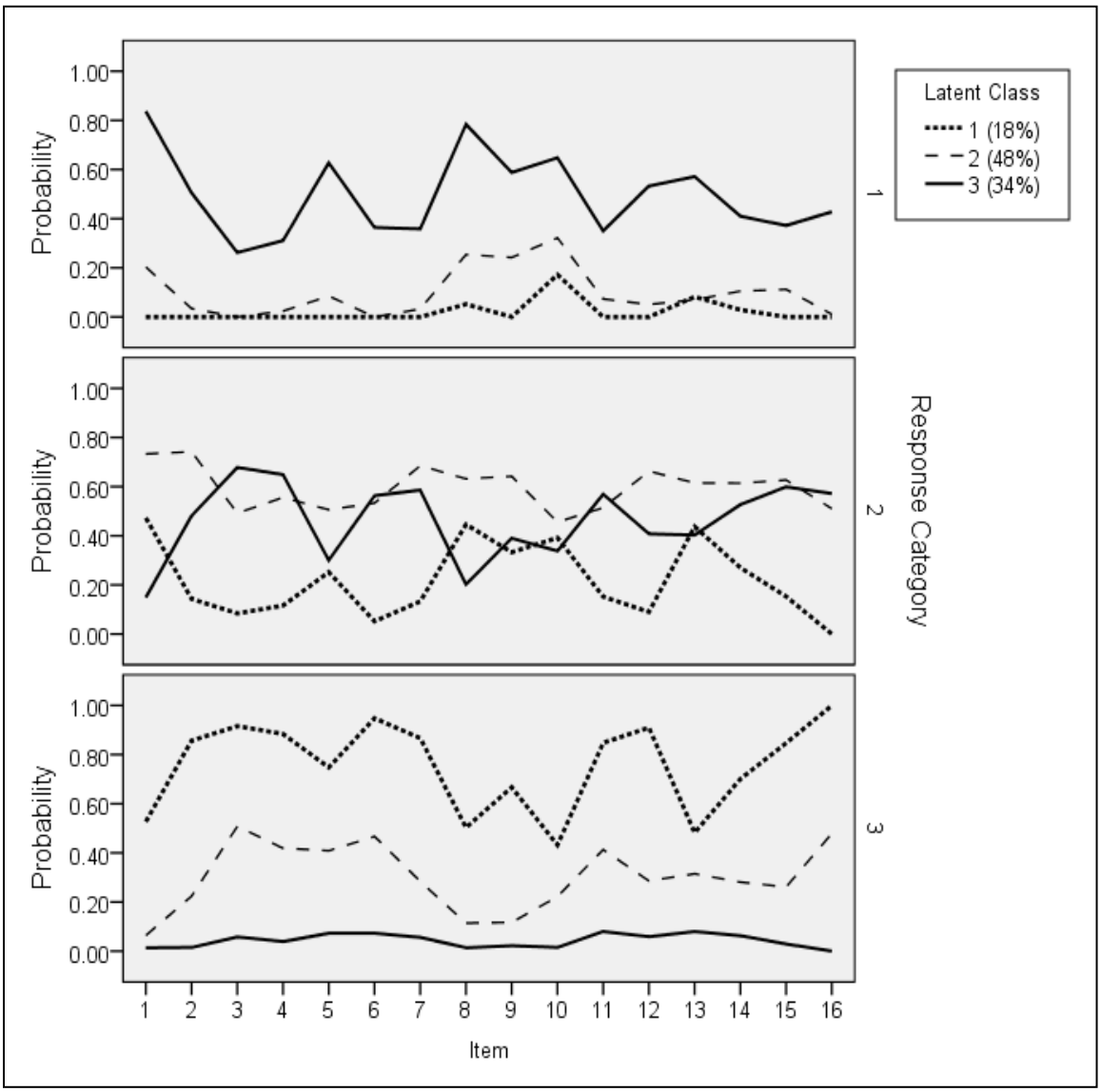


Figure 12. LCA time 1: Probability of item endorsement by latent class

Based on these results, Class 1 appears to be the high severity class ("*Clinical symptom severity*") as this class has a high probability of endorsing each externalizing symptom with the highest frequency. Class 2 is the moderate severity class ("*Subclinical symptom severity*") with highest probabilities of endorsing each symptom with the middle response category, indicating moderate frequency. Class 3 is the low severity class ("*Nonclinical symptom severity*") as this class has the highest probabilities of endorsing each symptom with the lowest response category, indicating a lack or limited occurrence of symptoms.

GRM

Parameter Estimates

The estimated factor loadings $(\hat{\lambda}_i)$ and item thresholds $(\hat{\delta}_{ik})$, as well as their associated standard errors, resulting from the GRM analysis with data from time 1 are found in Table 13. To aid interpretation, the right three columns contain the traditional IRT item parameter transformations computed from Mplus results: item discrimination (a) and severity threshold (b) parameters.

Discrimination parameters were all positive and ranged from 0.64 to 2.20. These parameters indicate the strength of the relationship between an item and the measured construct and how well the item is able to discriminate between youth above and below the item threshold. This is seen by the slope on the item characteristic curve (ICC) at the value of the item threshold. For example, the ICC for the highest response category of item 1 and item 16 is shown in Figure 13. The steeper slope seen for item 16 indicates

that this item is better able to differentiate between youth above and below the item threshold compared to item 1. The threshold value is the level of the latent trait where the ICC is at a probability of .50. As expected with ordered category responses, the thresholds were ordered with the items' first thresholds being negative and ranging from −0.35 to −1.65 and the items' second thresholds being positive and ranging from 0.23 to 1.60 (see Table 13).

Table 13. Item Parameter Estimates from GRM at Time 1: Mplus GRM Analysis Results and Equivalent IRT Estimates (J = 204)

		Mplus Formulation			IRT Formulation		
Item		Loading (*SE*)	Thresh 1 (*SE*)	Thresh 2 (*SE*)	Disc	Severity 1	Severity 2
1	B01	2.28 (0.31)	−0.80 (0.26)	3.22 (0.39)	1.34	−0.35	1.41
2	A02	2.60 (0.37)	−2.71 (0.38)	1.91 (0.32)	1.53	−1.04	0.73
3	B04	2.40 (0.35)	−3.95 (0.48)	0.55 (0.26)	1.41	−1.65	0.23
4	A03	2.22 (0.32)	−3.33 (0.40)	0.90 (0.26)	1.30	−1.50	0.41
5	A04	2.01 (0.28)	−1.68 (0.27)	0.99 (0.24)	1.18	−0.84	0.49
6	A05	2.82 (0.41)	−3.74 (0.49)	0.70 (0.29)	1.66	−1.33	0.25
7	B06	2.14 (0.30)	−2.97 (0.37)	1.34 (0.26)	1.26	−1.39	0.63
8	B07	1.77 (0.25)	−0.61 (0.22)	2.54 (0.30)	1.04	−0.34	1.44
9	A10	1.58 (0.23)	−1.10 (0.22)	2.09 (0.27)	0.93	−0.70	1.32
10	A11	1.08 (0.19)	−0.50 (0.18)	1.73 (0.22)	0.64	−0.46	1.60
11	B10	1.76 (0.25)	−2.48 (0.30)	0.75 (0.22)	1.04	−1.41	0.43
12	B11	2.50 (0.35)	−2.40 (0.35)	1.42 (0.29)	1.47	−0.96	0.57
13	A13	1.30 (0.20)	−1.57 (0.22)	1.36 (0.21)	0.76	−1.21	1.05
14	B12	1.42 (0.22)	−1.88 (0.25)	1.26 (0.22)	0.83	−1.32	0.89
15	A15	1.86 (0.27)	−2.31 (0.30)	1.39 (0.25)	1.10	−1.24	0.75
16	B15	3.74 (0.59)	−4.16 (0.63)	0.94 (0.37)	2.20	−1.11	0.25

Note: Thresh = Threshold; Disc = Discrimination.

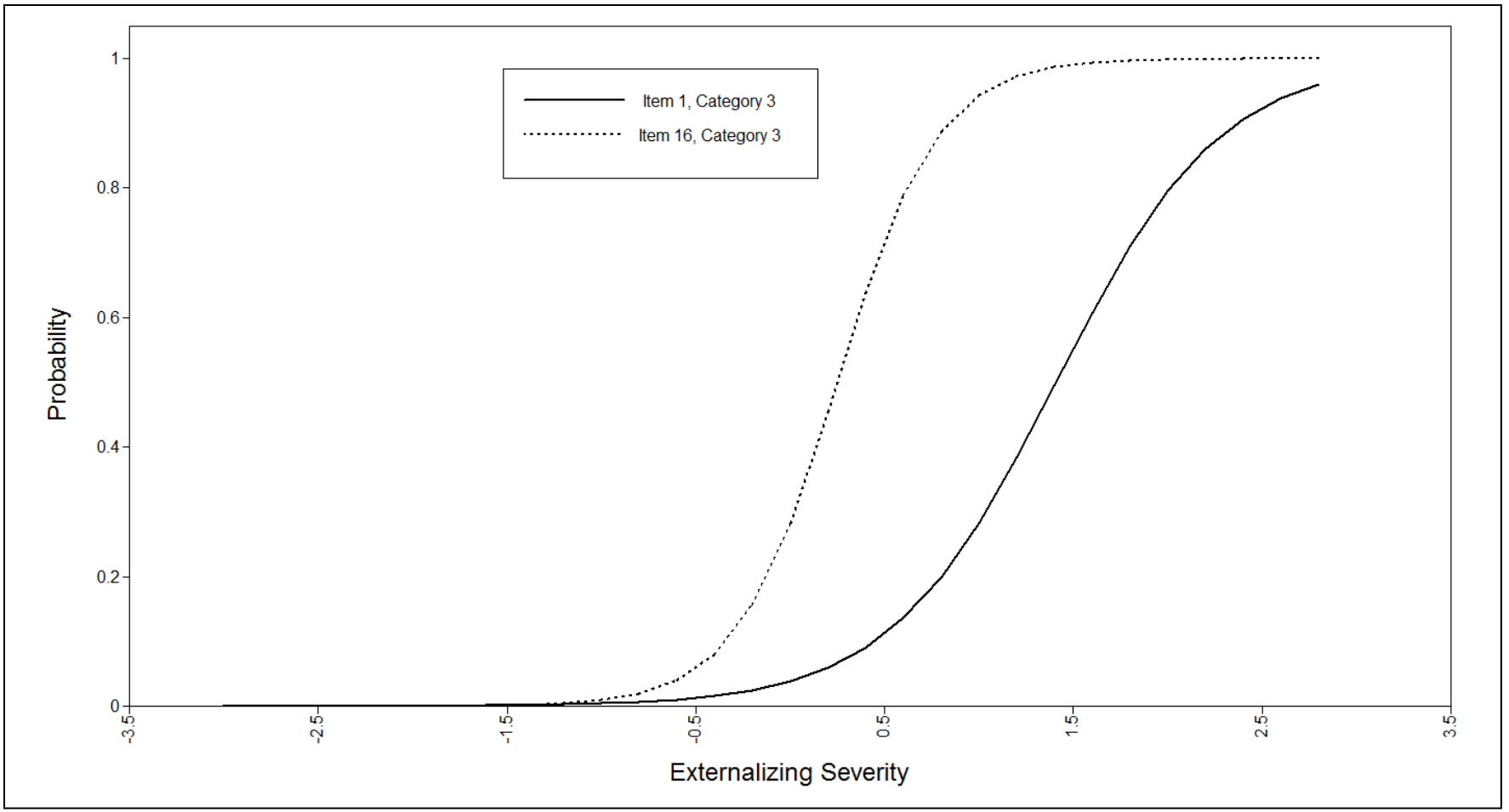


Figure 13. ICC curves for the highest response category for items 1 and 16.

Predicted Severity

With GRM analysis, Mplus produces predicted values of the person's externalizing symptom severity $(\tilde{\theta}_j)$. These values are in a logit scale where higher values indicate higher levels of externalizing symptom severity. Basic descriptives of these predicted values are found in Table 14. Externalizing severities ranged from −2.63 to 2.44, with approximately 72% of the sample having a predicted severity between −1.0 and 1.0 logits.

Table 14. Predicted Externalizing Symptom Severity at Time 1 (J = 204)

Statistic	Value	Percentile	Value
Mean	−0.01	10	−1.20
SD	0.96	20	−0.81
Median	0.05	30	−0.51
Minimum	−2.63	40	−0.21
Maximum	2.44	50	0.05
		60	0.26
		70	0.44
		80	0.74
		90	1.23

*on a logit scale

LTA

As described in Chapter 2, application of a LTA model proceeded in three steps. These are: 1) estimation of LCA model parameters separately at each time point; 2) exploration of transitions based on LCA results; and 3) estimation of LTA model parameters.

Step 1: LCA Model at Each Time Point

Procedures and results from LCA at time 1 (i.e., treatment start) were presented at the beginning of the current chapter. The results from time 2 (i.e., treatment end) are presented below.

Model selection. To select the number of latent classes for the final model, LCA models with an increasing number of classes were fit to the data at time 2. According to the BIC and LMR-LRT fit indices (see Table 15), the model with three latent classes provided the best fit at time 2. Thus, the model with three latent classes was selected as the final model at both time points.

Table 15. Fit Indices for Model Selection (Time 2)

# Latent Classes	# of Parameters	Treatment End	
		BIC[1]	LMR-LRT[2]
1	32	6134.93	
2	65	5365.71	939.36, $p < .001$
3	98	**4962.82**	**575.11, $p < .001$**
4	131	5037.19	100.56, $p = .566$

Note: BIC = Bayesian information criterion; LMR-LRT = Lo Mendell Rubin Likelihood Test
[1]Lowest number indicates best fit
[2]Significance indicates the model fits significantly better than a model with one fewer class

Item parameter estimates. The item threshold estimates ($\hat{\delta}_{ikg}$) resulting from application of the LCA model at time 2 are found in Table 16. Thresholds approaching an extreme were set to −15 or 15 by Mplus. Item thresholds correlated with thresholds from time 1 (see Table 11) at $r = .68$ for threshold 1 and at $r = .48$ for threshold 2. This indicates that there are some differences in the item thresholds across time, meaning the

measure may not be time invariant. As will be discussed later, this can be problematic for interpretation of youth change over time.

Class assignment. Participants were assigned to a latent class at time 2 based on their largest posterior probability. The resulting frequency of class assignment for the sample is displayed in Table 17. Class 2 was the largest with 99 youth (49%), followed by Class 1 with 62 youth (30%), and Class 3 was the smallest with 43 youth (21%). The average probability for these class assignments was .98 in each class. The relative entropy was .95.

Class Descriptions. Similar to the results from LCA at time 1, the probabilities of endorsement of the SFSS externalizing items were inspected within each class in order to describe the latent classes according to externalizing symptom severity, and to see if the classes had the same meanings as those found at time 1. These probabilities are depicted in Figure 14. For all items, the largest probability of endorsement at the highest response category (i.e., indicating the highest frequency/severity of symptoms) was found for Class 1 and the highest probability of endorsement at the lowest response category (i.e., indicating the lowest frequency/severity of symptoms) was found for Class 3. Based on these results, the latent classes at time 2 have similar meanings as for those from time 1: Class 1 appears to be the high severity class ("*Clinical symptom severity*"), Class 2 is the moderate severity class ("*Subclinical symptom severity*"), and Class 3 is the low severity class ("*Nonclinical symptom severity*").

Table 16. LCA Item Threshold Estimates at Time 2

	Threshold 1						Threshold 2					
	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 1		Latent Class 2		Latent Class 3	
Item	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
1	−2.20	0.44	−0.51	0.22	1.78	0.44	1.05	0.30	4.58	1.01	15.00	-
2	−3.34	0.72	−2.79	0.45	0.56	0.34	−0.19	0.27	2.70	0.43	3.72	1.01
3	−15.00	-	−4.58	1.01	0.25	0.32	−1.92	0.43	2.00	0.34	3.59	1.01
4	−4.09	1.01	−3.87	0.72	0.54	0.34	−0.94	0.30	1.83	0.32	3.95	1.28
5	−4.10	1.01	−2.23	0.36	1.05	0.36	−0.18	0.26	2.30	0.37	3.01	0.73
6	−4.09	1.01	−15.00	-	0.44	0.33	−1.37	0.35	2.83	0.49	15.00	-
7	−3.36	0.72	−3.35	0.59	0.50	0.34	−0.11	0.27	3.67	0.88	15.00	-
8	−1.67	0.37	−0.80	0.23	2.52	0.60	1.20	0.31	3.81	0.72	15.00	-
9	−2.18	0.43	−1.33	0.26	1.27	0.39	0.78	0.29	2.87	0.50	15.00	-
10	−1.78	0.39	−0.69	0.22	1.19	0.39	0.01	0.27	3.39	0.59	2.96	0.73
11	−3.37	0.72	−3.50	0.63	−0.02	0.32	−0.27	0.27	2.14	0.36	2.53	0.60
12	−3.46	0.81	−2.07	0.34	1.49	0.44	−0.49	0.28	3.01	0.52	3.66	1.01
13	−2.63	0.52	−2.67	0.47	0.55	0.33	0.25	0.27	2.51	0.40	3.68	1.01
14	−2.59	0.52	−1.99	0.32	−0.04	0.32	−0.03	0.27	2.30	0.37	3.64	1.01
15	−15.00	-	−2.39	0.39	0.35	0.33	−0.68	0.29	2.17	0.36	3.68	1.01
16	−4.09	1.02	−3.01	0.57	0.31	0.32	−1.30	0.34	1.85	0.31	3.68	1.01

dash (-) = not calculated because estimate was fixed

Table 17. Final Class Counts and Proportions (J = 204) at Time 2

Latent Class	Frequency	Proportion (%)
1	62	30
2	99	49
3	43	21

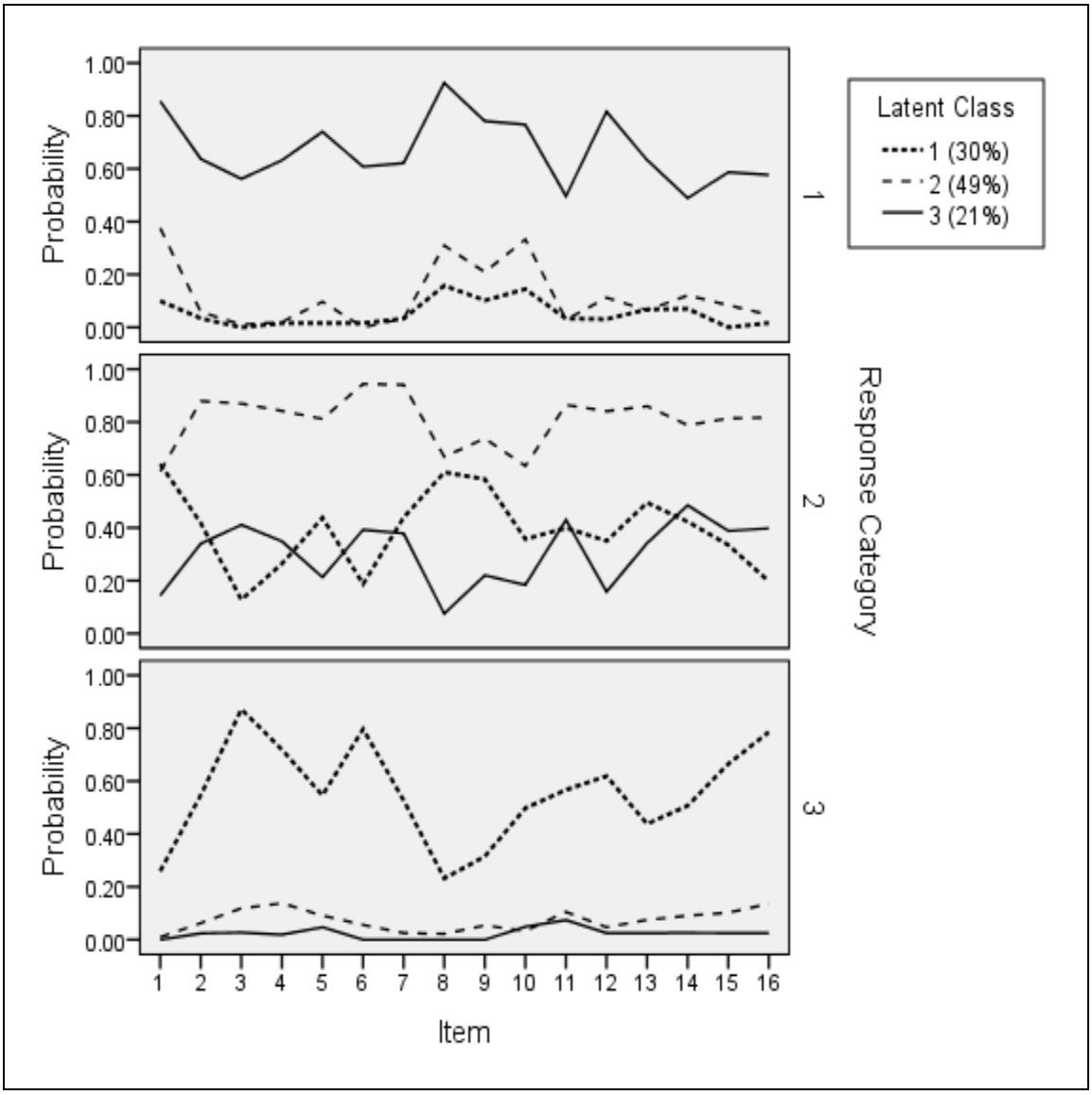


Figure 14. LCA time 2: Probability of item endorsement by latent class at time 2

Step 2: Transitions Based on LCA Model Results

Based on the LCA results at each time point, the movement of youth between latent classes was inspected across time. Results are depicted in Table 18. Numbers along the diagonal represent youth who remained in the same latent class across time. For example, 23 youth who were in the Clinical symptom severity class at treatment start were also in the Clinical class at treatment end. Youth located below the diagonal

changed latent class membership in a negative direction, indicating deterioration or worsening of symptoms. For example, 38 youths began in the Nonclinical class but were classified in the Subclinical class at treatment end, indicating deterioration. Youth above the diagonal showed positive change in latent class membership, moving from a class characterized by higher externalizing symptom severity to one with lower severity, thus showing improvement. For example, 10 youths who were in the Clinical class at the beginning of treatment were in the Subclinical class at treatment end.

Table 18. Latent Class Membership across Time Based on Cross-sectional results

		Time 2			
		Clinical	Subclinical	Nonclinical	Total
Time 1	Clinical	**23**	10	4	37
	Subclinical	32	**51**	15	98
	Nonclinical	7	38	**24**	69
	Total	62	99	43	204

Step 3: LTA Model Application

The assumption of measurement invariance is important to consider in the application of the LTA model as it has implications for the interpretation of transition probabilities. Measurement invariance assumes the equality of the parameters of the measurement model (i.e., the conditional item probabilities estimated for each class) across time. If full invariance holds, the conditional item probabilities are the same across time points and the transition probabilities have a straightforward interpretation. If invariance does not hold, change in latent class membership may be a result of the change in item parameters and not from true change in class membership. Given the latent class

solutions had the same number of latent classes at each time point and the profiles of the conditional item probabilities looked similar (i.e., the classes were interpreted the same across time), full measurement invariance is a plausible assumption in the current demonstration. Therefore, LTA proceeded under the assumption of full measurement invariance. For more discussion about measurement invariance within LTA, see Nylund (2007).

Two different approaches can be taken when fitting a LTA model assuming full measurement invariance. In the first approach, the class-specific item parameters obtained from LCA at the first time point are used as fixed values for subsequent time points. In the second approach, the estimation of item parameters occurs simultaneously for time 1 and time 2 but are constrained to be equal across time. These methods are referred to as *fixed* and *joint* estimation, respectively (Cho, Cohen, Kim, & Bottge, 2010). Both methods were used here.

Fixed estimation method. Using the fixed estimation method, item parameters at each time point were fixed at the values obtained with LCA at time 1. Thus, the conditional item probabilities at each time point were identical to those shown in Figure 9. The resulting latent transition probabilities using the fixed estimation method are found in Table 19. These probabilities reflect the probability of latent class membership at the end of treatment (time 2) based on latent class membership at the beginning of treatment (time 1) (the second term in Equation 3 from Chapter 2). For example, youth in the Clinical class at the beginning of treatment had a probability of .34 of still being in the Clinical class at the end of treatment, and a probability of 0.55 of moving to the Subclinical class.

Table 19. Latent Transition Probabilities from LTA with Fixed Estimation

		Time 2		
		Clinical	Subclinical	Nonclinical
Time 1	Clinical	0.34	0.55	0.11
	Subclinical	0.01	0.73	0.21
	Nonclinical	0.05	0.35	0.61

Youth were classified into latent classes based on their largest posterior probabilities. The final class counts and proportions for each time point are found in Table 20. The classification of youth at time 1 was identical to that found from LCA at time 1 (refer to Table 12). At time 2, the Subclinical class was the largest with 116 youth (57%), the Nonclinical class was next with 66 youth (32%), and the Clinical class was the smallest with 22 youth (11%).

Table 20. Final Class Counts and Proportions from LTA with Fixed Estimation

	Latent Class	Class Label	Frequency	Proportion (%)
Time 1	1	Clinical	37	18
	2	Subclinical	98	48
	3	Nonclinical	69	34
Time 2	1	Clinical	22	11
	2	Subclinical	116	57
	3	Nonclinical	66	32

Based on latent class assignment, the movement of youth between latent classes across time is depicted in Table 21. Similar to Table 18, the diagonal represents youth who did not change latent class membership from the start to the end of treatment. For example, 14 youths in the Clinical class at time 1 were still in the Clinical class at time 2. Youth whose symptom severity improved, as indicated by a change in latent class membership from a higher severity class to a lower severity class, are found above the

diagonal. Those who deteriorated are below the diagonal. In total, these results indicate 45 youths improved, 34 deteriorated, and 125 did not change in externalizing symptom severity through the course of treatment. The average probability for latent class assignment ranged from .82 to 1.00. The relative entropy was .90.

Table 21. Latent Class Membership from LTA with Fixed Estimation (N = 204)

		Time 2			
		Clinical	Subclinical	Nonclinical	Total
Time 1	Clinical	**13**	20	4	37
	Subclinical	6	**71**	21	98
	Nonclinical	3	25	**41**	69
	Total	22	116	66	204

Joint estimation method. With the joint estimation method, item parameters were estimated in the model but were constrained to be equal across time. Therefore, the probability of item endorsement within a latent class remained the same from time 1 to time 2. The estimated item thresholds ($\hat{\delta}_{ikg_t}$) resulting from the joint estimation method of LTA are found in Table 22. Mplus fixed thresholds approaching an extreme at −15 or 15. If full measurement invariance holds, it would be expected that the threshold estimates found here would be equivalent to those found with the fixed estimation method (see Table 11; LCA at time 1). However, they correlated only at r = .66 for Threshold 1 and at r = .70 for Threshold 2. Thus, parameter estimates appear to change over time, indicating a potential violation of the assumption of measurement invariance over time.

Table 22. LTA Item Threshold Estimates from Joint Estimation*

	Threshold 1						Threshold 2					
	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 1		Latent Class 2		Latent Class 3	
Item	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
1	−2.08	0.27	−0.41	0.17	1.83	0.33	1.01	0.19	4.64	0.97	4.62	1.23
2	−3.69	0.55	−2.23	0.28	0.49	0.24	−0.20	0.17	2.76	0.36	4.45	1.04
3	−15.00	-	−3.63	0.53	−0.20	0.23	−1.55	0.23	1.75	0.25	3.64	0.72
4	−4.43	0.82	−2.84	0.35	−0.01	0.23	−0.90	0.19	1.71	0.23	15.00	-
5	−3.06	0.40	−2.00	0.25	1.27	0.30	−0.36	0.17	1.72	0.24	4.18	1.01
6	−15.00	-	−2.98	0.38	0.05	0.23	−1.25	0.21	2.16	0.31	3.31	0.63
7	−3.40	0.47	−2.99	0.40	0.16	0.23	−0.26	0.17	2.88	0.40	3.28	0.59
8	−1.61	0.23	−0.51	0.18	1.68	0.31	0.95	0.19	4.00	0.73	4.43	1.01
9	−1.78	0.24	−1.26	0.20	1.14	0.27	0.69	0.18	2.78	0.36	15.00	-
10	−1.25	0.20	−0.57	0.17	0.97	0.26	0.34	0.17	3.05	0.44	3.30	0.59
11	−2.86	0.37	−3.69	0.65	0.08	0.23	−0.44	0.17	1.63	0.23	3.00	0.52
12	−3.90	0.62	−1.85	0.24	1.00	0.27	−0.48	0.18	2.81	0.39	3.00	0.52
13	−2.64	0.33	−2.27	0.30	0.68	0.24	0.28	0.17	2.03	0.26	3.35	0.63
14	−2.52	0.32	−2.16	0.29	0.23	0.23	0.01	0.17	2.08	0.27	3.31	0.60
15	−3.22	0.46	−2.17	0.28	0.17	0.23	−0.40	0.18	2.27	0.29	3.65	0.72
16	−4.31	0.73	−3.13	0.48	0.41	0.23	−1.30	0.22	1.73	0.24	15.00	-

*Parameters are equivalent at both time points; dash (-) = not calculated because estimate was fixed

Based on the results obtained from the joint estimation method, the probabilities of item endorsement were inspected within each class. These probabilities can be found in Figure 15. As expected, the results are similar to those found previously. For all items, the largest probability of endorsement at the highest response category (i.e., indicating the highest frequency/severity of symptoms) was found for Class 1, and the highest probability of endorsement at the lowest response category (i.e., indicating the lowest frequency/severity of symptoms) was found for Class 3. Based on these results, the classes have the same meaning assigned previously: Class 1 is the high severity class ("*Clinical symptom severity*"), Class 2 is the moderate severity class ("*Subclinical symptom severity*"), and Class 3 is the low severity class ("*Nonclinical symptom severity*").

The transition probabilities (the second term in Equation 3 from Chapter 2) resulting from the joint estimation method are found in Table 23. These are interpreted the same as in Table 19. For example, those in the Subclinical class at time 1 had a probability of .69 of remaining in the Subclinical class at time 2, indicating they did not improve or deteriorate. These results show a similar pattern to those found with the fixed estimation method, with a few exceptions. Based on the joint estimation method, those in the Clinical class at time 1 had the highest probability of remaining in the Clinical class and a total probability of showing improvement (i.e., moving to the Subclinical or Nonclinical class) of .51. However, with the fixed estimation method results, those in the Clinical class at time 1 had the highest probability of moving to the Subclinical class at time 2 and a total probability of showing symptom improvement of .64. Additionally, those in the Nonclinical group at time 1 had a total probability of deterioration (i.e.,

moving to a Subclinical or Clinical class) of .39 based on the fixed estimation method and a probability of .50 based on the joint estimation method.

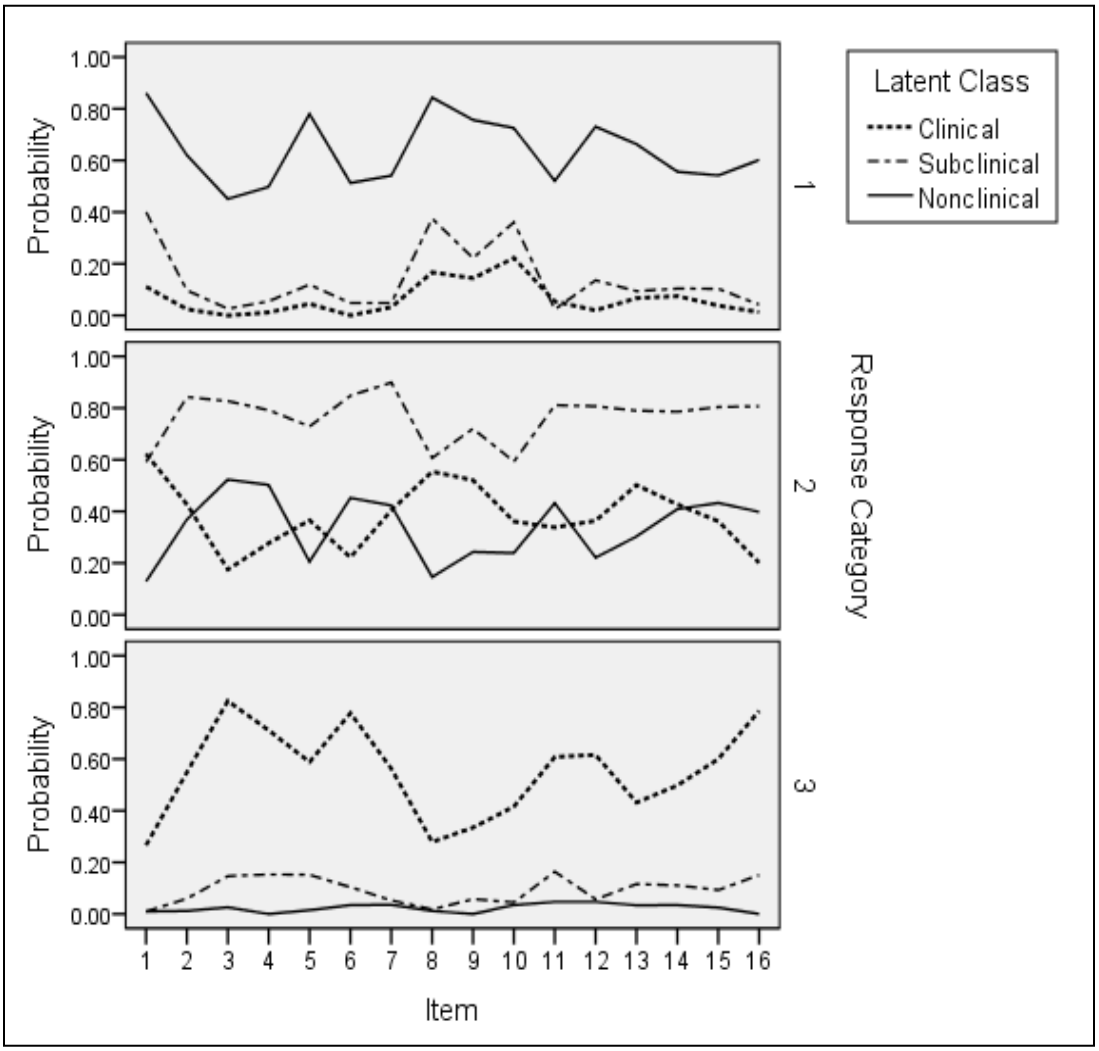


Figure 15. Class-specific probabilities of item endorsement based on joint estimation method

Table 23. Latent Transition Probabilities from LTA with Joint Estimation

		Time 2		
		Clinical	Subclinical	Nonclinical
Time 1	Clinical	0.49	0.43	0.08
	Subclinical	0.13	0.69	0.18
	Nonclinical	0.10	0.40	0.50

Youth were assigned to latent classes based on their largest posterior probabilities. The final class counts and proportions at each time point resulting from the joint estimation method are found in Table 24. At time 1, the largest class was the Clinical class with 97 youth (48%), followed by the Subclinical class with 62 youth (30%), and the smallest class was the Nonclinical class with 45 youth (22%). At time 2, the largest class was the Subclinical class with 103 youth (51%), followed by the Clinical class with 60 youth (29%), and the smallest class was the Nonclinical class with 41 youth (20%).

Table 24. Final Class Counts and Proportions from LTA with Joint Estimation

	Latent Class	Class Label	Frequency	Proportion (%)
Time 1	1	Clinical	97	48
	2	Subclinical	62	30
	3	Nonclinical	45	22
Time 2	1	Clinical	60	29
	2	Subclinical	103	51
	3	Nonclinical	41	20

The most likely classification of youth transitions between latent classes is depicted in Table 25. Similar to Tables 18 and 20, those on the diagonal did not change latent classes from time 1 to time 2, those above the diagonal showed latent class movement indicating symptom improvement, and those below the diagonal showed latent

class movement indicating symptom worsening or deterioration. According to this, 61 youth improved (30%), 113 remained the same (55%), and 30 deteriorated (15%). The average probability for latent class assignment ranged from .87 to 1.00. The relative entropy was .93.

Table 25. Latent Class Membership from LTA with Joint Estimation (J = 204)

		Time 2			
		Clinical	Subclinical	Nonclinical	Total
Time 1	Clinical	**48**	42	7	97
	Subclinical	7	**43**	12	62
	Nonclinical	5	18	**22**	45
	Total	60	103	41	204

In comparison to the results from the fixed estimation method, the largest difference appears when comparing the Clinical and Subclinical classes at time 1. With the fixed estimation method, 18% of the sample was classified in the Clinical class at time 1 and 48% were classified in the Subclinical class. However, with the joint estimation method, 48% were classified in the Clinical class at time 1 and 30% were classified in the Subclinical class. The difference in classification from these methods at the first time point is depicted in Table 26. The youth who were consistently classified across the methods are those along the diagonal. With LTA, 84 youth (41%) would be classified differently based on the estimation method used.

When comparing the movement in latent classes between time points in terms of whether youth improved (i.e., transitioned from a higher severity class to a lower severity class), deteriorated (i.e., transitioned from a lower severity class to a higher severity

class), or remained the same (i.e., did not change latent class), similar discrepancies occur. Results are found in Table 27. For a total of 70 youth (34%), change in severity would be described differently depending on which estimation method was used. For example, 28 youth who did not change latent classes with the fixed estimation method, showed a change in latent class membership indicating symptom improvement with the joint estimation method. As a whole, these differences may further indicate the measure does not maintain invariance over time, making the interpretation of transitions difficult. However, for the demonstration purposes of the current dissertation to compare statistical output across models, analyses proceeded under the assumption of full invariance. If the SFSS externalizing subscale is used in this manner in the future, more work will be needed to ensure appropriate interpretations are possible.

Table 26. Time 1 Class Assignment Comparisons: Fixed and Joint Estimation

		JOINT (time 1)			
		Clinical	Subclinical	Nonclinical	Total
FIXED (time1)	Clinical	**37**	0	0	37
	Subclinical	60	**38**	0	96
	Nonclinical	0	24	**45**	69
	Total	97	62	45	204

Table 27. Latent Class Transitions: Fixed and Joint Estimation

		JOINT			
		Improvement	No Change	Deterioration	Total
FIXED	Improvement	**33**	12	0	45
	No Change	28	**84**	13	125
	Deterioration	0	17	**17**	34
	Total	61	113	30	204

Longitudinal GRM

Longitudinal Invariance

As discussed in Chapter 2, assessing longitudinal invariance is important for the interpretation of results from longitudinal GRM analysis. If longitudinal invariance does not hold, change over time may not be attributable to a true change in symptom severity. To investigate longitudinal invariance, DIF by time was assessed using the UCI and OCI procedures described and used in Chapter 3 for the psychometric analysis of the SFSS Externalizing subscale. Here, the grouping variable used to differentiate the focal group from the reference group was time (i.e., time 1 vs. time 2).

OCI DIF results. Results of OCI DIF procedures are found in Table 28. The GMH statistic indicated potential DIF for item 6 ($MH\chi^2 = 6.31$, $p = .04$) and item 11 ($MH\chi^2 = 9.28$, $p = .01$). To investigate the strength of DIF, the probability for person at time 1 indicating a different response category for an item compared to a person at time 2 was calculated when persons were matched across time by externalizing severity scores. For items 6 and 11, these probabilities were .55 and .57 respectively. Thus, when matched for overall Externalizing SFSS score, the probability that a caregiver chose a different response category for these items at time 2 compared to time 1 was .56 on average.

Results of LR analyses indicated potential uniform DIF for item 3 ($\Delta G^2 = 6.31$, $p < .05$) and potential non−uniform DIF for item 11 ($\Delta G^2 = 9.28$, $p < .05$). As introduced in Chapter 3, Zumbo (1999) proposed calculating two measures of magnitude by looking at the difference between the two reduced models in terms of their generalizing

coefficients of determination and the coefficients of determination rescaled by their maximum values. These values are in Table 29, all of which fall well below the proposed cutoff value of 0.13 (Zumbo, 1999).

Table 28. Results of OCI DIF Analyses for SFSS Externalizing Items by Time

Item and Description	GMH	Non−Uniform LR	Uniform LR
1: Throw things when mad	0.97	1.54	0.76
2: Get in trouble	1.86	0.88	0.29
3: Disobey adults	0.23	**5.37**[B]	0.69
4: Interrupt others	0.15	0.30	0.29
5: Lie to get things	5.44	3.02	0.04
6: Hard to control temper	**6.31**[A]	0.01	0.87
7: Not get along fam/friend	2.85	0.01	1.68
8: Threaten/bully others	0.97	0.10	0.31
9: Can't wait turn	1.42	0.10	0.01
10: Hang with troubled peers	5.51	2.53	**5.11**[B]
11: Can't pay attention	**9.28**[A]	2.38	1.54
12: Gets into fights	1.25	0.03	0.93
13: Loses things	4.12	0.10	0.26
14: Can't sit still	1.78	0.05	0.16
15: Annoy others on purpose	3.08	0.14	1.66
16: Argues with adults	1.55	2.36	0.08

GMH = Generalized Mantel−Haenszel statistic; bolded if significant at $p < .05$
LR = Logistic Regression; bolded if Chi square of difference in deviance statistics ($df = 1$) sig at $p < .05$.
[A] Probability of 0.75 or less of focal group responding different
[B] Negligible effect size = 0.05

Table 29. Effect Sizes for DIF items by LR Method

Item	Change in R^2	Change in Max R^2
3	0.01	0.01
10	0.01	0.01

R^2 = coefficient of determination; Max = rescaled to maximum value
Note: Values over 0.13 indicate significant DIF (Zumbo, 1999)

UCI DIF results. The results of applying a GRM separately to data at time 1 and time 2 are found in Table 28. Mplus results include parameter estimates for factor

loadings ($\hat{\lambda}_{it}$) as well as the item thresholds ($\hat{\delta}_{ikt}$) and their associated standard errors. Items 1 and 2 were used as anchor items to ensure that all estimates were on a common scale. Therefore, the item loading and thresholds for these items at time 2 were fixed at the values found at time 1 (see Table 13). The equivalent traditional IRT parameters were also calculated from Mplus item parameter estimates (see Table 31).

Similar to the scatterplots introduced in Chapter 3, scatterplots of SFSS Externalizing item parameters by time are found in Figures 16-18; Figure 16 depicts the first thresholds, Figure 17 the second thresholds, and Figure 18 the factor loadings. The solid line on each graph is the 45-degree reference line where items without DIF are expected to fall. Confidence intervals of approximately two standard errors for each parameter are drawn for each item.

According to the scatterplot for the first threshold (Figure 16), item 3 (‘*disobeys adults*’) and item 13 (‘*loses things’*) may have potential DIF since the confidence intervals fall slightly short of crossing the reference line. Additionally, results for the second threshold indicate the confidence intervals for items 3, 4, 5, 6, 7, 8, 11, 12, and 13 fail to cross the reference line (see Figure 17). In fact, these items fall above the line. This suggests potential DIF where the second threshold is consistently larger at time 2 compared to time 1 for these items. As seen when comparing item severities for the second threshold (i.e., IRT formulation in Table 31), endorsement of these items at or above the second threshold may indicate a higher level of symptom severity at time 2 compared to time 1.

Table 30. Estimates for Factor Loadings and Item Thresholds from GRM at each Time

Item		Time 1						Time 2					
		Loading	*SE*	Threshold 1		Threshold 2		Loading	*SE*	Threshold 1		Threshold 2	
				Threshold	*SE*	Threshold	*SE*			Threshold	*SE*	Threshold	*SE*
1	B01	2.28	0.31	−0.80	0.26	3.22	0.39	**2.28**	-	−0.80	-	**3.22**	-
2	A02	2.60	0.37	−2.71	0.38	1.91	0.32	**2.60**	-	−2.71	-	**1.91**	-
3	B04	2.40	0.35	−3.95	0.48	0.55	0.26	4.78	0.75	−5.86	0.89	1.78	0.38
4	A03	2.22	0.32	−3.33	0.40	0.90	0.26	2.79	0.37	−3.42	0.41	1.52	0.26
5	A04	2.01	0.28	−1.68	0.27	0.99	0.24	2.29	0.30	−2.31	0.29	1.95	0.26
6	A05	2.82	0.41	−3.74	0.49	0.70	0.29	4.39	0.65	−5.21	0.74	2.36	0.41
7	B06	2.14	0.30	−2.97	0.37	1.34	0.26	3.30	0.46	−3.71	0.48	3.00	0.41
8	B07	1.77	0.25	−0.61	0.22	2.54	0.30	2.07	0.30	−0.79	0.21	3.41	0.38
9	A10	1.58	0.23	−1.10	0.22	2.09	0.27	1.86	0.67	−1.39	0.22	2.63	0.30
10	A11	1.08	0.19	−0.50	0.18	1.73	0.22	1.70	0.25	−0.88	0.19	1.96	0.24
11	B10	1.76	0.25	−2.48	0.30	0.75	0.22	2.07	0.29	−3.02	0.35	1.67	0.24
12	B11	2.50	0.35	−2.40	0.35	1.42	0.29	2.94	0.39	−2.43	0.33	2.29	0.32
13	A13	1.30	0.20	−1.57	0.22	1.36	0.21	2.02	0.28	−2.37	0.29	2.20	0.27
14	B12	1.42	0.22	−1.88	0.25	1.26	0.22	1.73	0.25	−2.19	0.26	1.79	0.24
15	A15	1.86	0.27	−2.31	0.30	1.39	0.25	2.42	0.32	−2.87	0.34	1.68	0.26
16	B15	3.74	0.59	−4.16	0.63	0.94	0.37	3.03	0.41	−3.59	0.45	1.42	0.27

Note: Bolded values are fixed; dash (-) = not calculated because estimate was fixed

Table 31. Traditional IRT Location and Discrimination Parameters

Item		Time 1			Time 2		
		Discrimination	Severity (1)	Severity (2)	Discrimination	Severity (1)	Severity (2)
1	B01	1.34	−0.35	1.41	1.34	−0.35	1.41
2	A02	1.53	−1.04	0.73	1.53	−1.04	0.73
3	B04	1.41	−1.65	0.23	2.81	−1.22	0.37
4	A03	1.30	−1.50	0.41	1.64	−1.23	0.54
5	A04	1.18	−0.84	0.49	1.35	−1.01	0.85
6	A05	1.66	−1.33	0.25	2.58	−1.19	0.54
7	B06	1.26	−1.39	0.63	1.94	−1.12	0.91
8	B07	1.04	−0.34	1.44	1.22	−0.38	1.65
9	A10	0.93	−0.70	1.32	1.09	−0.75	1.42
10	A11	0.64	−0.46	1.60	1.00	−0.52	1.15
11	B10	1.04	−1.41	0.43	1.22	−1.46	0.81
12	B11	1.47	−0.96	0.57	1.73	−0.83	0.78
13	A13	0.76	−1.21	1.05	1.19	−1.17	1.09
14	B12	0.83	−1.32	0.89	1.01	−1.27	1.04
15	A15	1.10	−1.24	0.75	1.42	−1.18	0.69
16	B15	2.20	−1.11	0.25	1.78	−1.18	0.47

Note: Severity refers to IRT location parameters; (1) refers to the first threshold, (2) refers to the second threshold

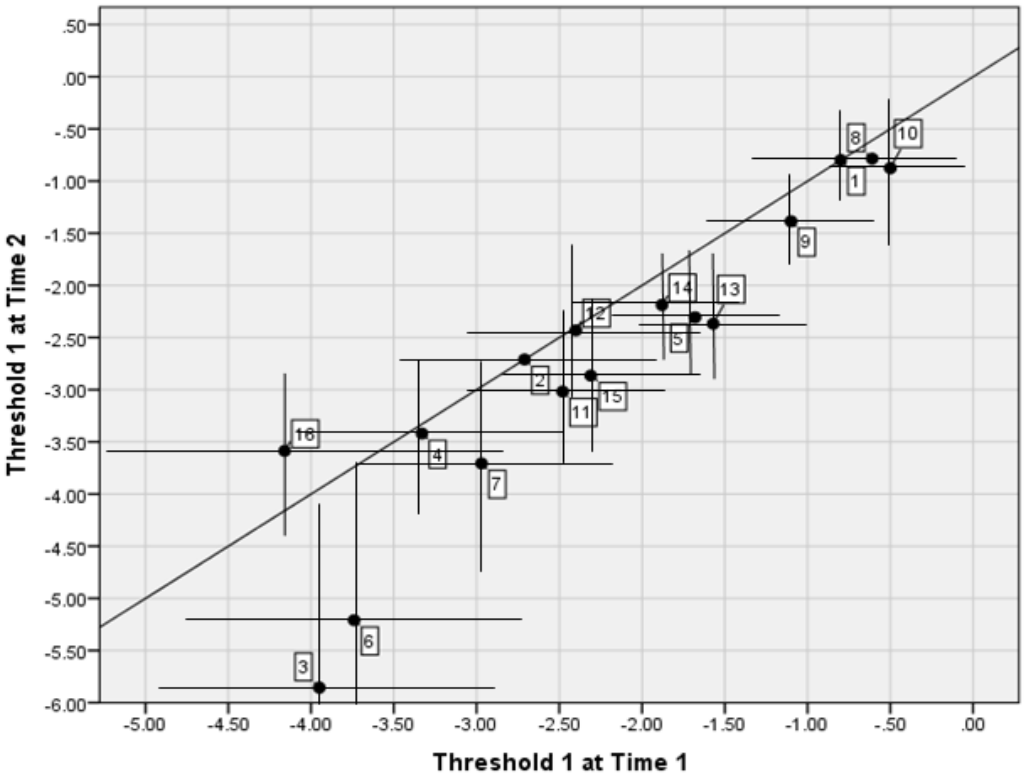


Figure 16. UCI scatterplot: Comparison of the first item threshold for items by time.
Note: Confidence intervals drawn around +/− 2*SE*.

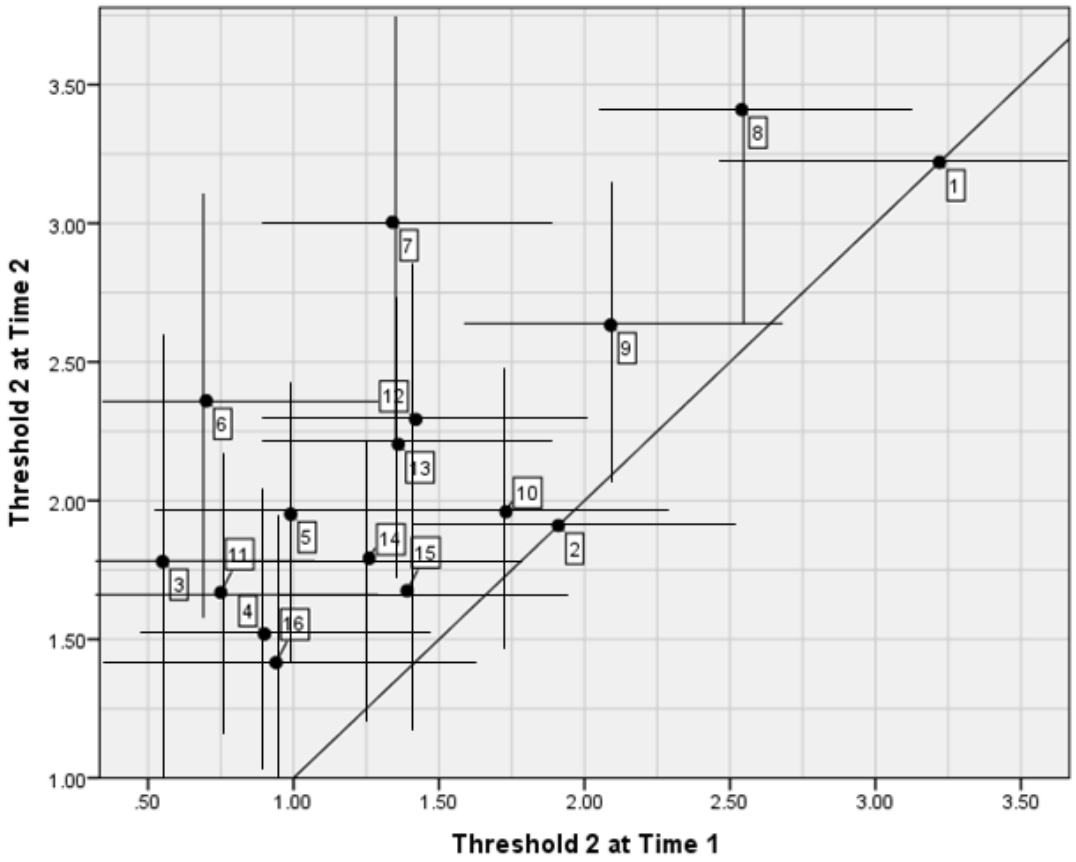


Figure 17. UCI scatterplot: Comparison of the second item threshold for items by time.
Note: Confidence intervals drawn around +/− 2*SE*.

The scatterplot comparing item factor loadings across time is found in Figure 18. Results indicate potential DIF for items 3, 6, 7, 10, and 13 as their confidence intervals fail to cross the reference line. This may mean these items are better able to differentiate respondents with high vs. low symptom severity at time 2 compared to time 1 (see discrimination values in Table 31). Unfortunately, with this UCI approach for detecting DIF, a lack of guidelines for determining what constitutes significant DIF in this approach makes it difficult to determine whether any of the differences found for either threshold or the factor loadings indicate a significant difference.

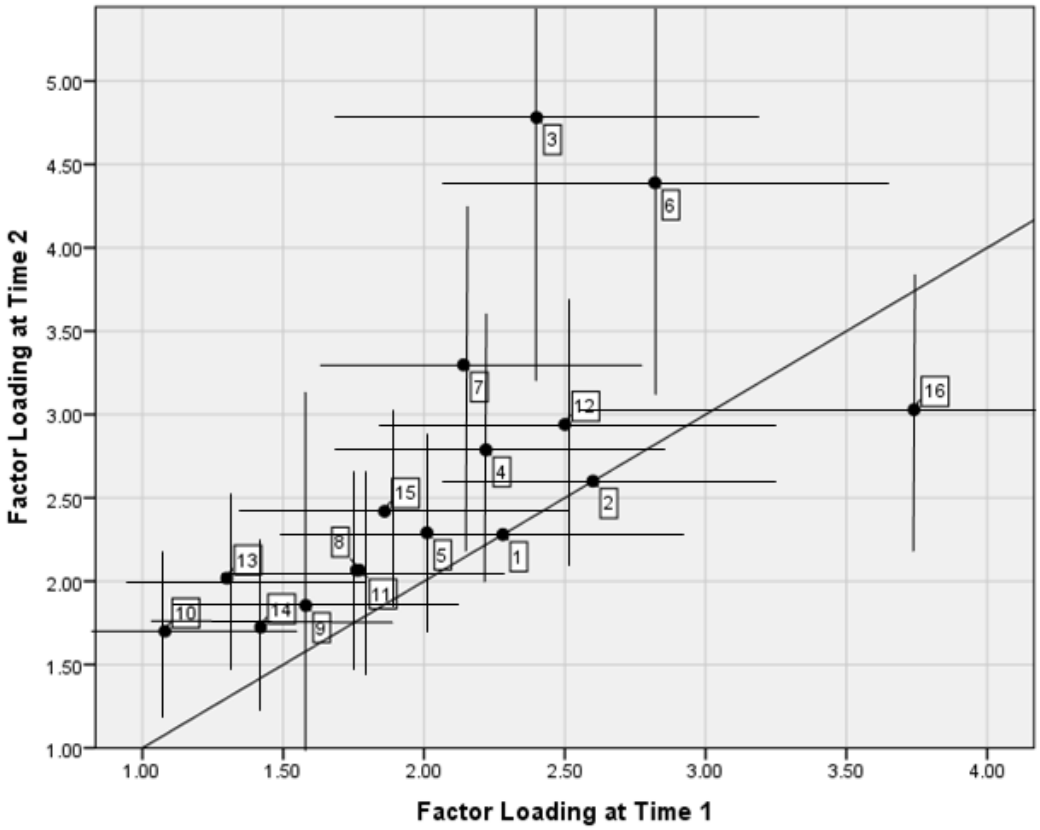


Figure 18. UCI scatterplot: Comparison of item factor loadings by time.
Note: Confidence intervals drawn around +/− 2*SE*.

Although OCI approaches did not indicate the possibility of much (if any) DIF, the UCI approach depicted a different picture. UCI results suggest that the statistical properties of the measure items (i.e., the parameters) changed over time. In other words, they are not time invariant. This is called item parameter drift (Goldstein, 1983) and can also seen when comparing item parameter estimates from time 1 and time 2 by calculating correlations. Here, factor loadings correlated at r = .63, the first thresholds correlated at r = 92, and the second thresholds correlated at r = 72. These results indicate possible invariance, particularly in the factor loadings and second thresholds, which is troublesome as it makes change in externalizing severity over time potentially

uninterpretable. Under those circumstances, using this measure for longitudinal IRT purposes would likely cause any statistical inferences to be biased. However, because the purpose of this application is to provide a demonstration and not to make conclusions concerning the longitudinal qualities of the SFSS externalizing subscale, longitudinal GRM analysis were conducted under the assumption of full measurement invariance.

Longitudinal GRM Application

Similar to the LTA model, the longitudinal GRM model can be estimated in two ways under the assumption of full measurement invariance: with fixed or joint estimation. Both methods were used here.

Fixed estimation method. Using the fixed estimation method, the item parameters (i.e., factor loadings and thresholds) were fixed for both time points at the values obtained in the GRM analysis for time point 1 (Table 13). As output, Mplus provides predicted values for youth externalizing symptom severity at both time 1 $(\tilde{\theta}_{j1})$ and time 2 $(\tilde{\theta}_{j2})$. Descriptives of these predicted values are found in Table 32. On a whole, the results were similar for both time points and the majority of youth had predicted severity between −1 and 1 logits with some youth having more extreme values. The covariance $(\hat{\sigma}_{12})$ between the latent variable at time 1 and time 2 was 0.51, indicating small positive relationship. This means that youth with higher scores at time 1 generally also have higher scores at time 2. This is expected; youth that begin treatment with the highest symptom severity would likely have higher severity compared to those who began with low symptom severity, even though they may have shown an overall decrease in severity.

However, because youth may increase or decrease in severity, it is also expected that the relationship between time 1 and time 2 scores is moderate.

Table 32. Predicted Symptom Severity based on Fixed Estimation (J = 204)*

	Time 1	Time 2
Mean	−0.01	−0.18
SD	0.96	0.89
Median	−0.15	−0.18
Minimum	−2.63	−2.50
Maximum	2.44	2.07
Percentiles		
10	−1.20	−1.39
20	−0.81	−0.83
30	−0.51	−0.46
40	−0.21	−0.30
50	0.05	−0.18
60	0.26	−0.04
70	0.44	0.21
80	0.74	0.55
90	1.23	0.95

*Values on a logit scale

The total change in severity was calculated by subtracting time 1 from time 2 ($\tilde{\theta}_{j2} - \tilde{\theta}_{j1}$). Thus, a negative number indicates a decrease in symptom severity (i.e., improvement) from time 1 to time 2, and a positive number indicates an increase in symptom severity (i.e., deterioration). Descriptives of these change scores can be found in Table 33. The amount of change in symptom severity ranged from −3.75 to 3.18 with more than half the sample showing an overall decrease in severity over time (i.e., negative change scores). However, these changes may not be statistically meaningful, as measurement error has not been accounted for.

Table 33. Change in Symptom Severity based on Fixed Estimation (J = 204)*

Statistic	Value	Percentile	Value
Mean	−0.17	10	−1.31
SD	0.89	20	−0.81
Median	−0.15	30	−0.57
Minimum	−3.75	40	−0.32
Maximum	3.18	50	−0.15
		60	0.03
		70	0.23
		80	0.42
		90	0.94

*values on a logit scale

To consider measurement error for youth severity estimates, an index of minimum detectable change (MDC) was calculated for each youth. An MDC represents the smallest change in scores from one measurement instance to the next that likely reflects true change rather than chance and measurement error alone (Schmitt & DeFabio, 2004). Typically, the MDC is calculated based on the standard error of the measurement (*SEM*) for the measure, resulting in a single MDC for all persons. However, because standard errors from IRT analyses can differ for each youth depending on their trait score, an MDC was calculated for each youth based on their unique standard errors from time 1 and time 2 scores[2]. The level of certainty represented by the MDC is determined by the respective z-score that is used in calculating it. Here, a 95% confidence interval was used. Thus, the MDC represents, with 95% certainty, the amount of change in symptom severity that represents real change for each youth. After this was calculated, the amount of change in each youth's predicted value was compared to the MDC to gain a clearer picture of where true change occurred as opposed to change potentially due to chance or

[2] MDC calculated as $1.96 * \sqrt{(SE_{time1})^2 + (SE_{time2})^2}$ for each youth

measurement error. Table 34 shows the number of youth whose change in their severity surpassed the MDC, thus likely indicating true trait change. Of the total 204 youth in the sample, 51 showed significant improvement (25%), 27 showed significant deterioration (13%), and 126 showed no change in their externalizing symptom severity (62%).

Table 34. Youth's Severity Change from Fixed Estimation Results*

	Frequency	% Total	Average Change (*SD*)
Significant improvement	51	25.0	−1.28 (0.58)
No change	126	61.8	−0.02 (0.38)
Significant deterioration	27	13.2	1.24 (0.52)

*Change classified based on MDC

Joint estimation method. In the joint estimation method, item parameters were estimated in the model but were constrained to be equal across time. The resulting estimated factor loadings ($\hat{\lambda}_{it}$) and thresholds ($\hat{\delta}_{ikt}$), as well as their associated standard errors, are found in Table 35. The equivalent parameters in traditional IRT formulation are also included. Note that one set of parameters is given, as they are identical at each time point. The results from the joint estimation method correlate highly with those found with the fixed estimation method for factor loadings (r = .91), first thresholds (r = .98), and second thresholds (r = .97). These high correlations may be interpreted as evidence of invariance across time; however, this conclusion is in contrast to evidence suggested by DIF analysis with the UCI approach discussed previously. In fact, these correlations may remain high even with significant (and systematic) item drift. However, given the lack of guidelines from the UCI approach for determining what constitutes significant DIF, these high correlations may also indicate that the differences detected

from the UCI DIF procedure were not significant. Therefore, it remains unclear whether a violation of measurement invariance has occurred. Future work is needed to further investigate measurement invariance of the SFSS externalizing subscale to determine whether any items display DIF.

Table 35. Mplus Longitudinal GRM Analysis Results and Equivalent IRT Estimates using Joint Estimation (J = 204)*

		Mplus Formulation			IRT Formulation		
Item		Loading (*SE*)	Thresh 1 (*SE*)	Thresh 2 (*SE*)	Disc	Severity 1	Severity 2
1	B01	2.08 (0.21)	−0.72 (0.18)	3.34 (0.29)	1.22	−0.35	1.61
2	A02	2.43 (0.24)	−2.70 (0.27)	2.13 (0.24)	1.43	−1.11	0.88
3	B04	3.12 (0.32)	−4.45 (0.43)	1.13 (0.26)	1.84	−1.43	0.36
4	A03	2.42 (0.24)	−3.32 (0.31)	1.29 (0.22)	1.42	−1.37	0.53
5	A04	2.04 (0.20)	−1.90 (0.21)	1.50 (0.20)	1.20	−0.93	0.74
6	A05	3.26 (0.34)	−4.16 (0.42)	1.49 (0.27)	1.92	−1.28	0.46
7	B06	2.46 (0.25)	−3.14 (0.30)	2.07 (0.25)	1.45	−1.27	0.84
8	B07	1.82 (0.19)	−0.63 (0.17)	3.98 (0.25)	1.07	−0.35	1.63
9	A10	1.63 (0.17)	−1.18 (0.17)	2.39 (0.21)	0.96	−0.73	1.47
10	A11	1.27 (0.15)	−0.63 (0.14)	1.86 (0.17)	0.75	−0.49	1.46
11	B10	1.83 (0.19)	−2.67 (0.23)	1.24 (0.18)	1.07	−1.46	0.68
12	B11	2.57 (0.25)	−2.32 (0.26)	1.90 (0.24)	1.51	−0.90	0.74
13	A13	1.55 (0.16)	−1.86 (0.19)	1.77 (0.18)	0.91	−1.21	1.15
14	B12	1.51 (0.16)	−1.98 (0.19)	1.56 (0.18)	0.89	−1.31	1.04
15	A15	2.03 (0.20)	−2.50 (0.24)	1.60 (0.20)	1.20	−1.23	0.79
16	B15	3.21 (0.33)	−3.73 (0.39)	1.30 (0.27)	1.89	−1.16	0.41

*Parameters are equivalent at both time points
Note: Thresh = Threshold; Disc = Discrimination

Descriptives of the predicted values for youth externalizing symptom severity at both time 1 ($\tilde{\theta}_{j1}$) and time 2 ($\tilde{\theta}_{j2}$) resulting from the joint estimation method are found in Table 36. On a whole, the results were similar for both time points. The majority of youth had predicted severity between −1 and 1 logits with some youth having more

extreme values. The covariance ($\hat{\sigma}_{12}$) between the latent variable at time 1 and time 2 was 0.49, indicating small a positive relationship. Similar to the fixed estimation method, this result is expected. This indicates that youth with higher severity scores at time 1 tended to have higher severity scores at time 2.

Table 36. Predicted Symptom Severity based on Joint Estimation (J = 204)*

	Time 1	Time 2[a]
Mean	0.09	−0.09
SD	0.99	0.93
Median	−0.17	−0.10
Minimum	−2.57	−2.45
Maximum	2.54	2.17
Percentiles		
10	−1.16	−1.39
20	−0.77	−0.78
30	−0.45	−0.43
40	−0.16	−0.22
50	0.17	−0.10
60	0.42	0.10
70	0.59	0.36
80	0.89	0.67
90	1.36	1.09

*Values on a logit scale; [a]Time 2 scores represent score at time 1 and change at time 2

Similar to the fixed estimation method, change scores for the youth were calculated by subtracting their predicted severity at time 1 from time 2 ($\tilde{\theta}_{j2} - \tilde{\theta}_{j1}$). Therefore a negative score indicates a decrease in symptom severity (i.e., improvement) and a positive score indicates an increase in symptom severity (i.e., deterioration). Description of these scores is found in Table 37. Change scores ranged from −3.86 to 3.27 with more than 50% of the youth showing an overall decrease in their severity estimate, before measurement error is taken into account.

Table 37. Change in Symptom Severity based on Joint Estimation (J = 204)*

Statistic	Value	Percentile	Value
Mean	−0.18	10	−1.36
SD	0.92	20	−0.87
Median	−0.15	30	−0.61
Minimum	−3.86	40	−0.32
Maximum	3.27	50	−0.15
		60	0.02
		70	0.24
		80	0.49
		90	0.93

*values on a logit scale

Similar to the fixed estimation method, youth were categorized based on whether they showed significant improvement (i.e., negative change > MDC), significant deterioration (i.e., positive change > MDC, or no change (i.e., change < MDC). The results are found in Table 38. Based on the joint estimation method, 55 youth showed significant improvement (27%), 30 showed significant deterioration (15%), and 119 showed no change in externalizing symptom severity (58%).

Table 38. Youth's Severity Change from Long GRM-Joint Estimation Results*

	Frequency	% Total	Average Change (*SD*)
Significant improvement	55	27.0	−1.29 (0.60)
No change	119	58.3	−0.02 (0.37)
Significant deterioration	30	14.7	1.22 (0.55)

*Change classified based on MDC

The predicted values of youth severity were very similar between the fixed and joint estimation method, with correlations of .99 (p < .001) between the predicted severities at the first time point, the second time point, as well as between the change scores. However, a few differences are noticeable when comparing the classification of

change. When youth were classified based on whether the amount of change in their predicted severity surpassed the MDC, the results differ slightly based on which estimation method was used. This can be seen in Table 39 where only youth found in the diagonal were classified the same across both methods. There was consistency between the methods for a large majority of the sample (frequency = 197; 97%). For example, 119 youth who showed significant improvement according to results from the fixed estimation method also showed significant improvement according to results from the joint estimation method. There were 7 youth (3%) whose change in severity would be classified differently based on which method is used. For example, three youth who deteriorated according to the fixed estimation results showed no change according to the joint estimation results. However, overall results were very similar across estimation method.

Table 39. Change in Severity Estimates from Long GRM: Fixed and Joint Estimation*

		JOINT (Change)			
		Significant Improvement	No Change	Significant Deterioration	Total
FIXED (Change)	Significant Improvement	**51**	4	0	55
	No Change	0	**119**	0	119
	Significant Deterioration	0	3	**27**	30
	Total	51	126	27	204

*Change classified based on MDC

Youth Severity: Comparison of LTA and Longitudinal GRM Results

The application of the LTA model and longitudinal GRM to the same set of data allows for the direct comparison of the results that would be used to provide clinical feedback. This type of comparison has yet to be published. Table 40 compares the results of the LTA and longitudinal GRM analysis at time 1 and 2 when fixed estimation was used, and Table 41 compares the results of LTA and longitudinal GRM analysis when joint estimation was used. Based on the fixed estimation method at time 1, the average predicted symptom severity was 1.38 logits for those in the Clinical class, 0.20 for those in the Subclinical class, and −1.04 for those classified in the Nonclinical class. The average severity significantly differed by latent class ($p < .001$) and displayed a pattern consistent with the qualitative labeling of the latent classes. Results were similar at each time point, and also across estimation method (see Tables 40 and 41 respectively).

Table 40. Comparison of Youth Output from Time 1 and 2 by Fixed Estimation

Time 1			Time 2		
Latent Class	Frequency	Mean Severity *	Latent Class	Frequency	Mean Severity*
Clinical	37	1.38	Clinical	22	1.26
Subclinical	98	0.20	Subclinical	116	0.10
Nonclinical	69	−1.04	Nonclinical	66	−1.14

* Means differ significantly by Latent Class, $p < .001$

Table 41. Comparison of Youth Output from Time 1 and 2 by Joint Estimation

Time 1			Time 2		
Latent Class	Frequency	Mean Severity *	Latent Class	Frequency	Mean Severity*
Clinical	97	0.90	Clinical	60	0.97
Subclinical	62	−0.22	Subclinical	103	−0.16
Nonclinical	45	−1.25	Nonclinical	41	−1.46

* Means differ significantly by Latent Class, $p < .001$

For a visual picture of the comparison between the results of LTA and longitudinal GRM analysis, each youth's predicted severity was plotted based on their latent class assignment at time 1 and 2 for both the fixed estimation method (Figure 19) and the joint estimation method (Figure 20). For example, Figure 19 shows the externalizing severity for each youth at time 1 and 2 from the fixed estimation method. Results are similar for both time points. The distributions of predicted values of severity appear to be ordered along the latent continuum, and are consistent with the labeling of the latent classes: youth in the clinical class have the highest severity, youth in the subclinical class have severity in the middle of the continuum, and youth in the nonclinical class have the lowest severity. Based on this, it appears that, while there are three distinct groups of youth with differing levels of externalizing severity, there is also some within-group variability. The same is observed when looking at the results from the joint estimation method (Figure 20).

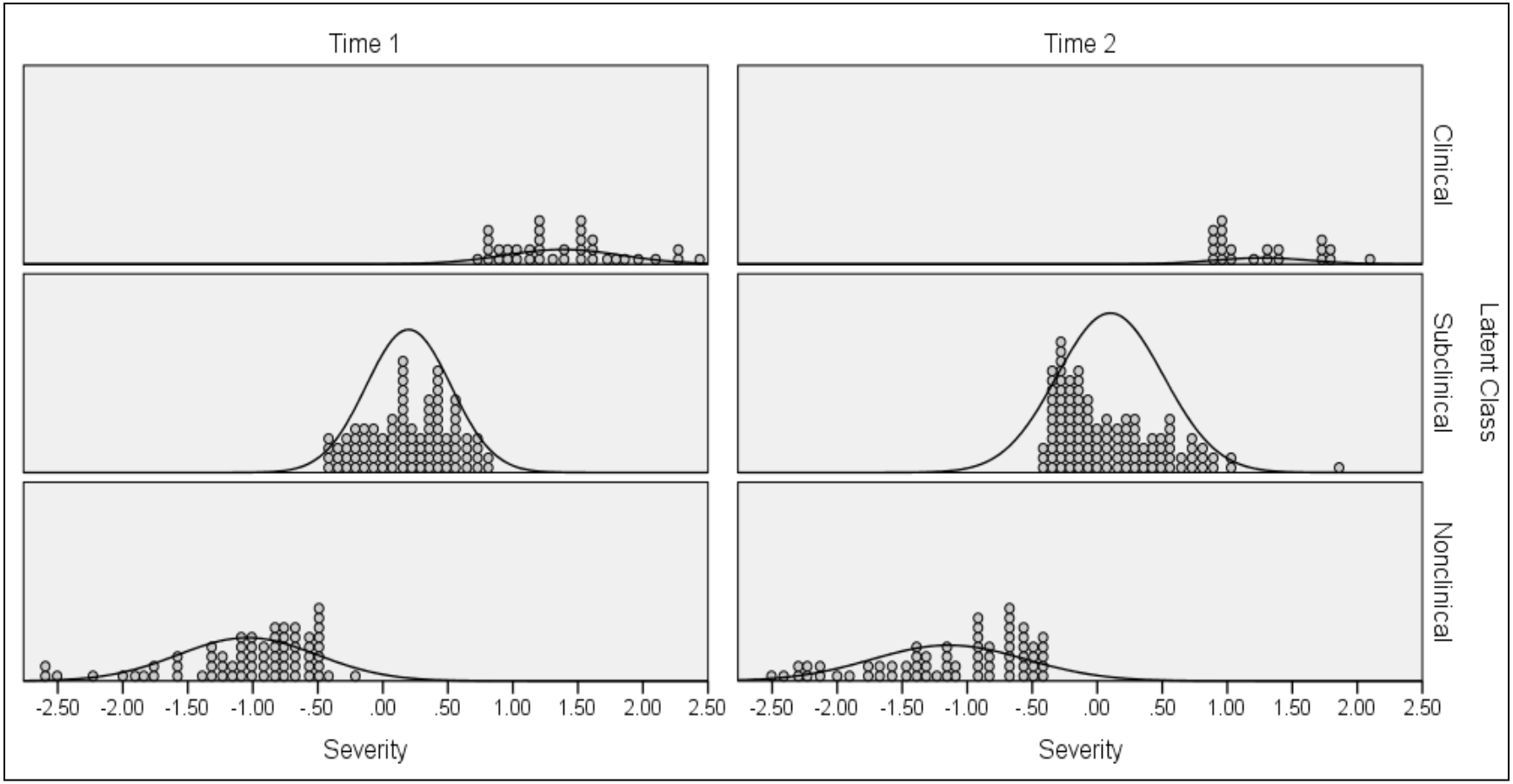


Figure 19. Fixed estimation method: youth predicted severity by latent class assignment.

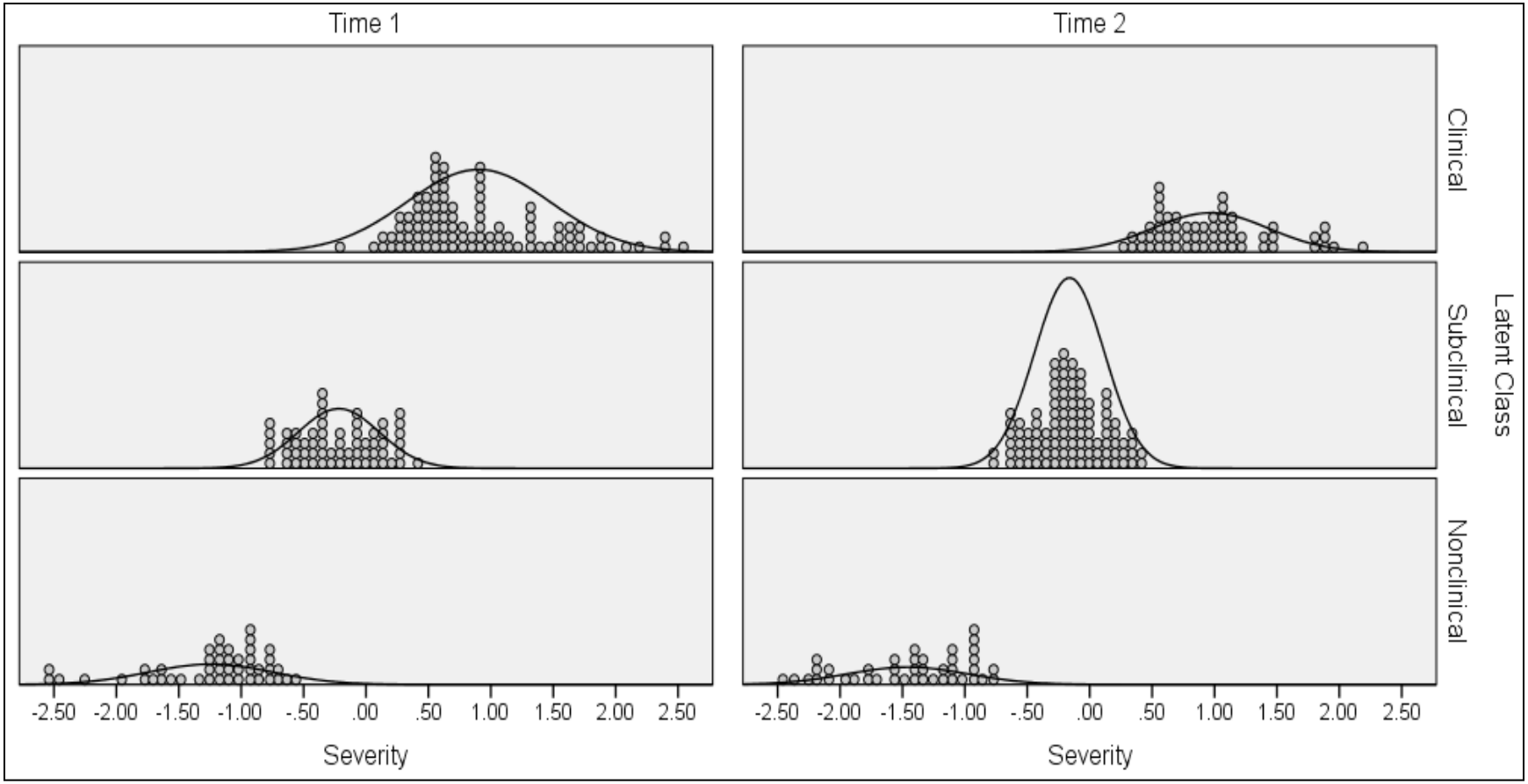


Figure 20. Joint estimation method: youth predicted severity by latent class assignment.

Table 42 and 43 compares youth change from time 1 to time 2 based on the results of longitudinal GRM analysis and LTA using both fixed and joint estimation methods respectively. After classifying youth according to whether they improved, deteriorated, or did not change, the longitudinal GRM analysis and LTA agreed on classification for 79% of the sample with fixed estimation and 77% of the sample with joint estimation. For example, as seen in Table 42, 38 youth who displayed significant improvement according to longitudinal GRM analysis (i.e., their externalizing severity decreased more than their index of MDC) using the fixed estimation method, also displayed improvement in the LTA (i.e., latent class movement from a higher severity class to a lower one) using the fixed estimation method.

Table 42. Comparison of Youth Change by Fixed Estimation

GRM Change	LTA Change			
	Improvement	No Change	Deterioration	Total
Significant improvement	**38**	13	0	51
No Change	7	**104**	15	126
Significant Deterioration	0	8	**19**	27
Total	45	125	34	204

Table 43. Comparison of Youth Change by Joint Estimation

GRM Change	LTA Change			
	Improvement	No Change	Deterioration	Total
Significant improvement	**45**	10	0	55
No Change	16	**93**	10	119
Significant Deterioration	0	10	**20**	30
Total	61	113	30	204

The similarity in results from LTA and longitudinal GRM analysis when youth were classified according to whether they improved, deteriorated, or did not change in their externalizing symptom severity can also be seen graphically in Figure 21.

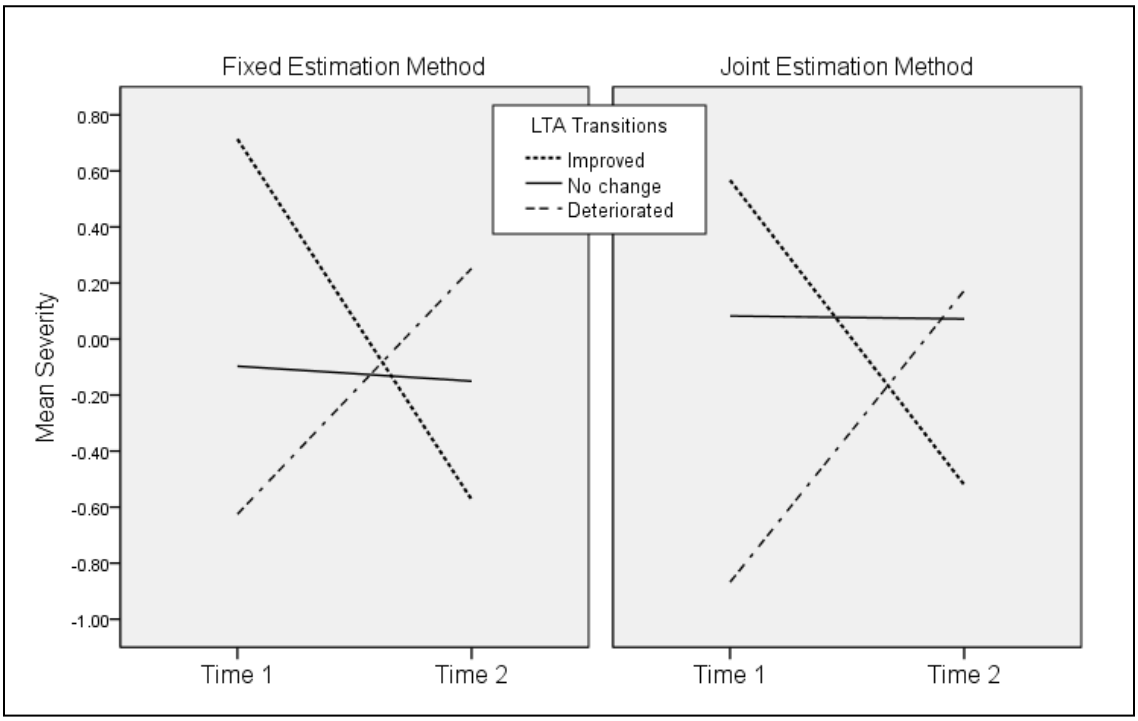


Figure 21. Mean severity based on latent transitions

As seen in Figure 21, the average severity from the longitudinal GRM analysis at each time point was plotted based on latent class assignment. The left panel depicts the results from using the fixed estimation method, and the right panel depicts the results from the joint estimation method. As can be seen, changes in average severity from time 1 to time 2 are consistent with the classification from LTA. For example, according to fixed estimation results (left panel), the dotted line with the steep and negative slope, showing a decrease in average severity from time 1 to time 2, belongs to the latent classes

that transitioned from a latent class with higher symptom severity to one of lower symptom severity. In other words, on average, youth represented by this line improved according to both LTA and longitudinal GRM analysis results.

Clinician Survey: Preferences for Feedback in Decision-Making

Although decision-makers can process information in any way they choose, they tend to process it (i.e., form their mental representation) in a way that is consistent with its presentation (e.g., Bettman & Kakkar, 1977; Chandra & Krovi, 1999). In other words, a categorical presentation will be naturally processed or interpreted in the mental representation in a categorical way; the information is not automatically transformed mentally into a different presentation format. Any transformation of information would require more cognitive effort. Thus, the cognitive load for using categorical information is lower when making a categorical decision, as no transformation is needed to apply the information. In this way, whether a clinical task is categorical or dimensional in nature can inform the presentation that limits the cognitive load.

As part of aim 3, to explore whether specific clinical tasks can be labeled as categorical or dimensional in nature, an informal survey (as described in Chapter 2) was completed by a total of nice practicing clinicians. A copy of this survey can be found in Appendix A. This informal survey asked clinicians about how useful they perceived specific presentations of information as feedback (i.e., categorical or dimensional) for clinical tasks. The goal of this survey was to determine whether clinicians indicate a preference for one presentation over another when making different clinical decisions.

From this, it is inferred that clinician preference for one presentation over when approaching a specific task represents a natural match, or a cognitive fit. Information processing theory suggests that humans have an automatic tendency to simplify the decision-making process, particularly in the face of high cognitive load. Thus, clinician preference for one presentation over another may indicate which presentation produces the lowest cognitive load for the decision.

The nine clinicians in this sample had been practicing for an average of 9.8 years (*SD* = 3.71, range = 5 – 17), primarily in outpatient settings, although all clinicians had some experience practicing in inpatient settings with two clinicians having more than 5 years experience. All nine clinicians served child and youth clients (i.e., under age 18) and five clinicians also served adults. Seven clinicians had Doctoral Degrees and two had Masters Degrees.

The results from this informal clinician survey are found in Table 44. Clinicians preferred the dimensional presentation to the categorical presentation when making decisions concerning whether clients are improving (F (1, 16) = 6.25, p = 0.02), whether clients are deteriorating (F (1, 16) = 7.22, p = 0.02), and whether the client is developing new symptoms (F (1, 16) = 5.49, p = 0.03). Additionally, preference for the dimensional presentation to the categorical presentation neared significance for making decisions about whether the treatment is working for the client (F (1, 16) = 4.38, p = 0.05). No other differences reached statistical significance.

It is important to note that this was an informal survey conducted with a limited sample size. Furthermore, an assumption was made that clinician preference indicates where cognitive fit occurs, which may not be true. However, this information provides a

starting point for carrying the empirical results in aim 2 into a more detailed illustration of cognitive fit for clinical decision-making. As will be mentioned in the discussion, more research is needed to better understand if clinical decisions can be categorized as categorical or dimensional in nature, and how they match (or do not match) with information presentations.

Table 44. Clinician Presentation Preferences based on Clinical Task (N = 9)

	Categorical Presentation		Dimensional Presentation		Difference	
	Mean	*SD*	Mean	*SD*	F	*p*
Psychopathology	4.33	1.32	3.33	1.23	2.77	0.12
Level of Care	3.22	1.20	3.44	1.01	0.18	0.68
Diagnosis	3.78	1.30	3.67	1.00	0.04	0.84
Medication	3.33	1.32	3.33	1.00	0.00	1.00
Treatment Working	3.56	1.33	4.56	0.53	4.38	**0.05**
Getting Better	3.56	1.23	4.67	0.50	6.25	**0.02**
Change Treatment Plan	2.89	1.05	3.67	0.71	3.38	0.09
Termination	3.44	1.24	3.56	0.88	0.05	0.83
Deteriorating	3.11	1.17	4.33	0.71	7.22	**0.02**
Session Frequency	2.78	1.20	3.33	0.87	1.27	0.30
Hospitalization	2.56	1.13	2.78	1.09	0.18	0.68
New Symptoms	2.78	1.09	3.78	0.67	5.49	**0.03**

The next chapter presents a discussion summarizing the findings and conclusions of this dissertation. Each of the four specific aims introduced in Chapter 1 are reviewed and discussed.

CHAPTER 5

Discussion

The overall goal of this dissertation was to add a novel perspective to the discussion concerning whether the structure of clinical latent variables is categorical or dimensional in nature. The conceptualization of these variables is important because, not only does it provide the foundation for how they are discussed and assessed by researchers and practitioners, it also potentially affects how clinical feedback is used by clinicians for making decisions throughout the mental health treatment process. The statistical method used to analyze clinical data makes an assumption about the underlying structure of the latent variable. This then affects the resulting statistical output that is used to present clinical information to clinicians to aid them in the clinical tasks that are central to the treatment process. This dissertation proposes that understanding how the presentation of information in feedback affects clinical decision-making provides a clinically useful means of deciding whether to analyze clinical data assuming a categorical or dimensional latent variable structure. To this end, there were four specific aims. These will each be discussed in turn.

Specific Aims

Aim 1: To apply a general model of decision-making to clinical decision-making and discuss the role of cognitive fit for determining the most effective presentation of clinical information in feedback.

The first aim, accomplished in Chapter 1, offered the concept of cognitive fit as the underlying theory supporting the idea that the efficiency and effectiveness of clinical decisions is related to the presentation of clinical information. Based on this concept, which is deeply rooted in information processing theory, clinical decisions are most efficient and effective when the information included in feedback used to make the decision is presented in a form that matches with the specific clinical task. Hence, information presented categorically is ideally matched with categorical tasks and information presented dimensionally is ideally matched with dimensional tasks. For example, a categorical approach may be most appropriate for decisions that differentiate unique groups (i.e., clinically severe vs. subclinical severe groups), and a dimensional approach may be most appropriate for decisions that involve individual differences in degree (i.e., numerical gradations of severity). In light of this, identifying the nature of the specific clinical tasks that are completed on an ongoing basis within clinical practice will allow for the deliberate choice of information presentation in feedback that will enhance clinician's decision-making and ultimately improve client outcomes.

Aim 2: To compare model output resulting from the application of statistical models to cross-sectional and longitudinal clinical data that assume different latent variable structures (i.e., categorical and dimensional).

The results presented in Chapter 4 serve as the foundation for this aim. Although statistical models assuming different latent variable structures have been statistically compared in terms of how well they fit to clinical data according to model fit indices (for example, see Krueger Markon et al., 2005; Walton et al., 2011), the output that would be used to provide clinicians with feedback about their individual clients has yet to be compared. Great consistency between categorical and dimensional approaches would provide evidence for the presence of an underlying dimension of severity (i.e., quantitative differences) while at the same time identifying unique groups along this continuum (i.e., qualitatively different groups). If this were the case, clinicians may receive a consistent message about their client (i.e., whether they are improving, deteriorating, or not changing), just presented differently. However, if there is a great dissimilarity between the results of these approaches, there is greater difficulty reconciling these perspectives and the feedback received by clinicians could be dramatically different. For example, a dimensional approach may yield a quantitative score relating to level of symptom severity, whereas a categorical approach with the same data may yield group membership to a class best described by a combination of present/absent externalizing diagnoses (e.g., CD, ODD, ADHD, etc.). In this case, the choice between approaches is more complicated as it not only affects the presentation of the information, but also the overall nature of the information provided.

Growing evidence suggests that latent classes resulting from clinical data are ordered along an underlying continuum of severity (e.g. see Lacourse et al., 2010; Nock et al., 2006). This was also found in the current application. Thus, one might expect the overall message in feedback about youth status and change in symptom severity to be similar regardless of the assumption made about the latent variable structure; however, this has yet to be examined. Therefore, it is important to examine the consistency and inconsistency in the predictions made about individuals (i.e., changes in trait levels or group membership) between the categorical and dimensional approaches. This may help elucidate the implications of deciding to take a categorical or dimensional approach to analyzing clinical data for clinical feedback.

As a whole, the results obtained from methods assuming a categorical latent variable structure (i.e., LCA and LTA) and methods assuming a dimensional latent variable structure (i.e., GRM analysis and longitudinal GRM analysis) were similar at an individual level. There was a large overlap between these methods in the resulting classification of youth's symptom severity and change over time. For example, a majority (77 – 79% from fixed and joint estimation methods, respectively) of youth who demonstrated a significant decrease in their externalizing severity estimate (i.e., they improved) according to results from longitudinal GRM analysis also demonstrated a change in latent class membership indicative of improvement (i.e., transitioning from a higher severity class to a lower one) from LTA. Thus, regardless of whether a categorical or dimensional approach was used, the general message provided in clinical feedback about youth status and change in externalizing severity would be similar.

Although the broad classification of youth was similar from each approach, the presentation of the information in feedback differs based on the assumption made in the analysis about the latent variable structure. This then affects the specific interpretation of the results. In a categorical approach (i.e., LCA and LTA), groups were differentiated yielding qualitative information about class membership, whereas in a dimensional approach (i.e., GRM analysis and longitudinal GRM analysis), predictions based on graded differences were made about individuals, yielding quantitative information about specific levels along a continuous scale. Hence, despite a similar overall message about youth change in externalizing severity, the latent variable structure assumed in the modeling approach affected both the presentation and interpretation of the results. In this case, the choice between a categorical or dimensional approach may be best approached based on the usefulness of the specific presentation of information provided. This relates to aim three of this dissertation, which will be further discussed in the next section.

It is important to note a limitation of these findings is that the current application assumed full measurement invariance over time. If the SFSS externalizing subscale is not invariant across time, interpretation of any change in externalizing severity may not be attributable to true change, making any inferences from these results biased. While assuming full invariance was appropriate for the purpose of the exploratory nature of this illustration, analyses indicating potential DIF as well as differences found between the results from the fixed and joint estimation methods suggest potential violation of the invariance assumption. Future work is needed to investigate the psychometric equivalence of the SFSS over time to ensure its use in longitudinal assessment is valid.

While aim 1 was accomplished with the discussion in Chapter 1, and aim 2 was accomplished through the empirical application in Chapters 2 and 4, the remaining aims of this dissertation have yet to be completed, as they are dependent on aims 1 and 2. Therefore, the next two aims are discussed based on the results and conclusions drawn from the first two.

Aim 3: To utilize case examples from aim two, as well as information from informal clinician surveys, to demonstrate how the concept of cognitive fit proposes that certain presentations of clinical information support more effective and efficient decision-making based on the specific clinical judgment or decision being made.

Although the results from models assuming different latent variable structures were similar, the output produced differed in presentation and interpretation. This directly affects the clinical feedback report and potentially its role in clinical decision-making. For example, Figure 22 represents a hypothetical feedback report that might be seen by a clinician based on the categorical approach. The information included in this figure reflects actual results for one youth ("John Smith", male, aged 16) included in the analysis sample. John's first completed externalizing SFSS measure was on May 31, 2007 where he was classified as being in the Clinical symptom severity class. Approximately four months later, another externalizing SFSS measure was completed that classified him in the Subclinical class. Based on his movement between latent classes from time 1 to time 2, John's externalizing symptom severity decreased during treatment, indicating improvement. Given the categorical nature of this feedback, there is no information concerning the amount of improvement that John made in this time.

These results simply show that John's pattern of item endorsement was more similar to those in the clinical class at time 1 and the subclinical class at time 2.

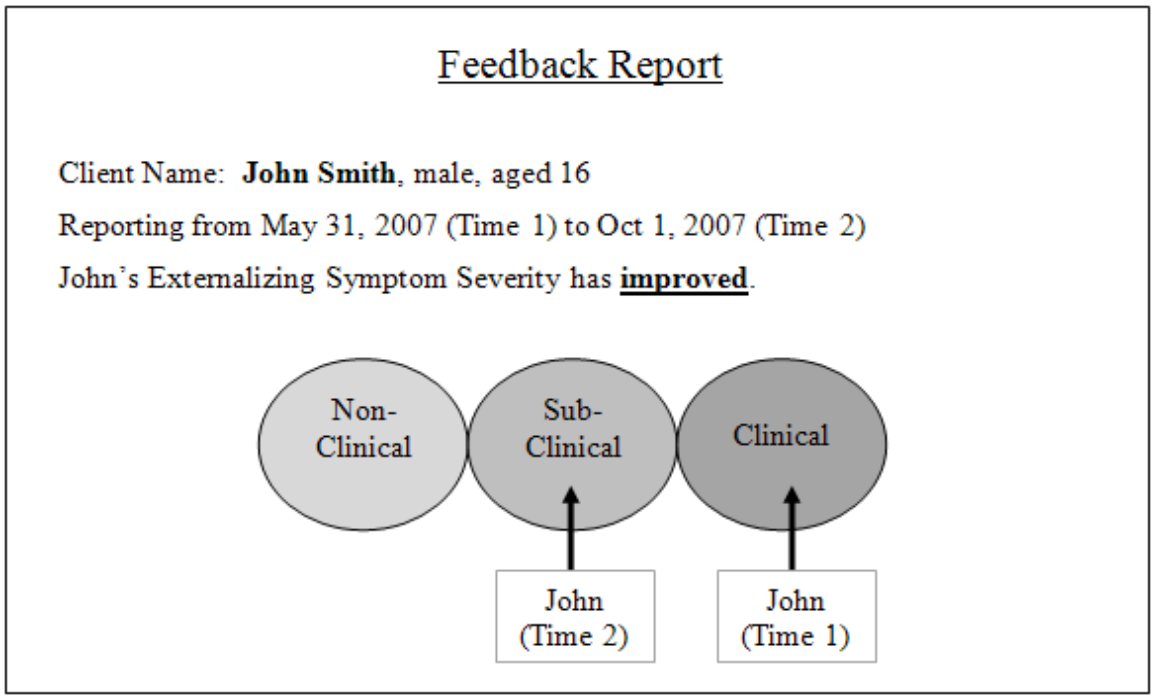


Figure 22. Hypothetical feedback report reflecting a categorical latent variable structure

When John's completed externalizing SFSS measure data were analyzed with a dimensional approach, the resulting hypothetical feedback report might look something like Figure 23. In this example, John had specific numbers assigned to his level of externalizing severity that can be located on the continuum (which also includes some reference numbers). At time 1, John's estimated severity was 1.96 logits but it decreased to 0.75 logits by time 2. Based on his calculated MDC, this change is deemed statistically significant and, similar to the first hypothetical feedback report, it is reported that John has improved in externalizing symptom severity over his four months in

treatment and by how much. In this presentation, the amount of change that John made is evident and a clinician is able to see how close to any cut-offs John now is (i.e., how far past the threshold into the medium category is John now?). However, this approach does not provide information about whether John is more or less similar to other groups of youth as the categorical approach does. It is important to note that, consistent with the use of the SFSS, this hypothetical example provides cut-points between low, medium, and high categories based purely on the percentiles from the psychometric evaluation. These categories do not identify clinical cut-points representing clinical or non-clinical severity levels. These percentiles are included to provide points of comparison to John's score. If these cut-points were actual indications of clinical, subclinical or non-clinical severity, it could be argued that the presentation in Figure 23 displays both categorical and dimensional information.

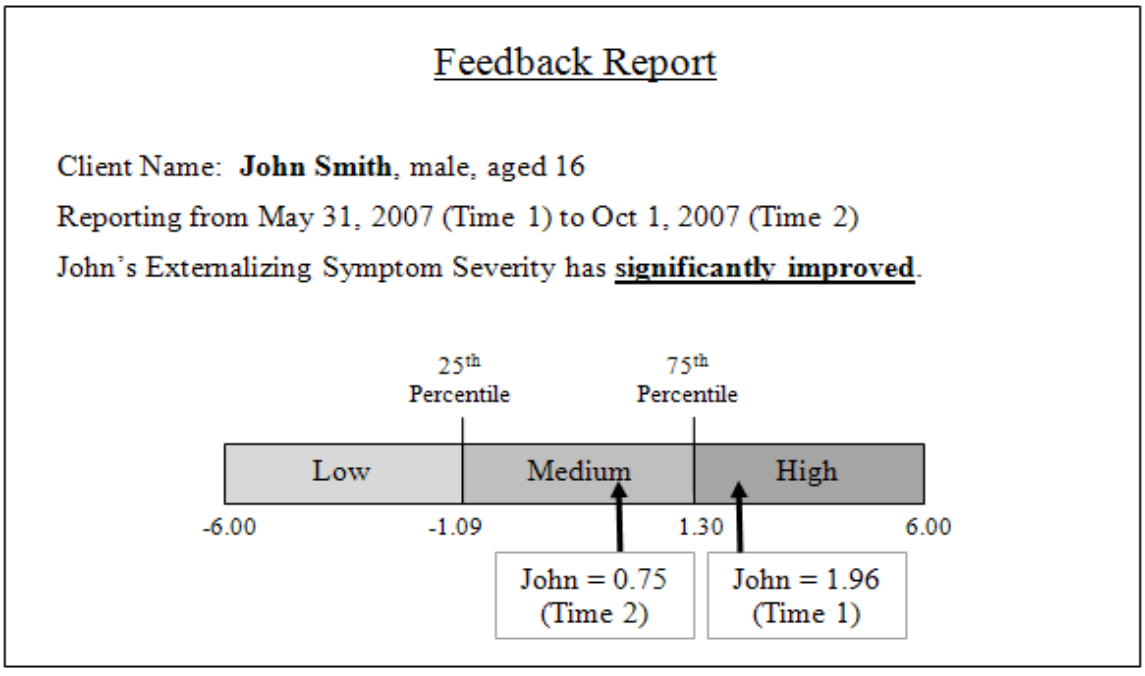


Figure 23. Hypothetical feedback report reflecting a dimensional latent variable structure

It is perhaps reassuring to see that the overall message about John's progress in treatment (i.e., he improved) is the same from both feedback reports. This indicates some consistency between statistical models. However, the presentation of the information differed greatly based on the model, and this difference may be important for clinical decision-making. For effective and efficient decision-making, cognitive fit theory suggests that clinicians review the feedback report with the presentation that matches with the clinical task. For example, if a clinical decision needs to be made based on group differences (e.g., perhaps whether John has symptoms similar to youth with ADHD), the first feedback report would be most useful for decision-making if youth with ADHD were classified in a specific class. However, if a clinical decision needs to be made based on graded changes (e.g., perhaps whether John's severity decreased at all or whether there is any indication that treatment is working), the later feedback report would be most useful. This hypothesis is novel and has never been tested.

In order to test the hypothesis of cognitive fit for the use of feedback in clinical decision-making, clinical decisions would need to be differentiated in a manner similar to tasks described in information processing literature (i.e., classifying tasks as spatial or symbolic) to match with spatial or symbolic information. However, currently there is no basis for labeling clinical decisions as being categorical or dimensional in order to match it with the presentation of information in feedback. This does not mean that decisions cannot be classified as such; it just means that they have not yet been looked at in this way. An exploratory approach was taken in the current dissertation by asking clinicians about their preference for one presentation over another for making specific clinical tasks. It was assumed that perhaps clinician preferences indicate the nature of the

decision or judgment (i.e., whether it is categorical or dimensional in nature) and its cognitive fit to the presentation of information.

Based on the informal clinical survey, clinicians preferred a dimensional presentation when making decisions about whether the treatment is working or whether a client is improving or deteriorating. These tasks were all similar in that they were all indicators of changes in symptom severity. This is consistent with the view that clinicians naturally think in a dimensional way when assessing change in symptom severity (e.g., Maser et al., 2009). If clinician preference indicates the cognitive fit between information presentation and the task, this suggests these types of decisions may be dimensional in nature. If this is the case, the theory of cognitive fit dictates that these decisions are most accurate and efficient when using the feedback report with the dimensional presentation of information (e.g., Figure 23), which provides a quantitative score representing differences in degree of symptom severity. According to this theory, using a feedback report with a categorical presentation potentially harms the decision-making process by increasing the clinician's cognitive load, thereby causing them to resort to using cognitive short cuts that result in decision error.

Clinicians did not clearly indicate preference for the categorical presentation for any decision; however, the list included in the survey was not comprehensive and the informal nature of this small study limits the ability to draw definitive conclusions. It simply served to illustrate this novel concept and suggest this as a potentially important area of future study. There may be clinical decisions that are most suited to information presented about distinct group differences where a categorical approach would be most appropriate. For example, if the categorical approach clearly indicated that the clinical

class contained all hospitalized children, receiving categorical information that a client's symptoms are most like other hospitalized children (i.e., the client is also in the clinical group), this may provide the best cognitive fit for deciding whether the client needs to be hospitalized or not.

At this point, some may advocate taking the best from both worlds by providing clinicians with both presentations. Why not provide as much information as possible; more is better, right? In fact, it may be rather intuitive to both researchers and clinicians to include as much detailed information as possible in feedback. In fact, several clinicians who took part in the informal survey commented that they would like a combination of both information presentations. However, research documents that including too much information can increase the cognitive load, thereby potentially increasing the reliance on heuristics (i.e., cognitive short-cuts), and may result in inaccurate decisions (Faust, 1984; Kliger & Kudryavtsev, 2010). Therefore, given the complex nature of clinical work and decisions being made, the strategy to include both presentations of information may backfire by inflating the clinicians' cognitive load. Clinicians would be forced to sort through larger amounts of information to discover what is relevant, and what is not relevant, or must mentally transform information into a format most useful for the clinical task. This may unintentionally harm the clinician's decision-making process.

Despite the potential risk of increasing the cognitive load by providing both presentations of information, it remains important to investigate this possibility. Currently, there is a wide range in the amount and kind of feedback that is provided to clinicians in research studies evaluating the effect of feedback on clinical outcomes. For

example, in the study conducted by Harmon et al. (2007), feedback consisted of a colored dot in the client's clinical chart. These dots, being either white, green, yellow, or red, indicated whether the client was functioning in the normal range, the client's rate of change was adequate or less than adequate, or if the client was not making expected progress based on a repeated measure of symptom severity. On the other hand, the feedback provided by Bickman et al., (2011) included the youth's numerical severity score along a dimension with reference points consistent with psychometric quartiles (similar to Figure 23) to identify whether youth's symptom severity is poor, fair, or good (color coded as red, orange, and green respectively). Furthermore, this feedback includes an indicator concerning the significance and direction of any change in score from the last measurement point (i.e., improvement, deterioration, or no change), a graph displaying these changes, and the youth's individual item responses. Although in these studies both Harmon et al. (2007) and Bickman et al., (2011) found significant effects of feedback on clinical outcomes, no research has investigated if and how the amount and format of the information influences these effects. Perhaps the effects in these studies would be stronger if more, less or different presentations of information were provided in the feedback reports because of a reduction in the cognitive load. Future research to identify the optimal type and amount of information to provide in feedback reports is warranted. The concept of cognitive fit and the effect of cognitive load on decision-making is one way explore this line of research.

Aim 4: To propose future research to investigate further the role of cognitive fit for informing how the assumed latent variable structure in a statistical model influences clinical feedback and decision-making.

While ample evidence in the mental health literature demonstrates that use of ongoing clinical feedback improves clinical outcomes (e.g., see Bickman et al., 2011; Lambert et al., 2005), how or why this occurs is unclear. Furthermore, little is known about whether the presentation of clinical data in feedback as categorical or dimensional information matters. Cognitive fit proposes that this presentation does matter and that, based on the specific clinical decision being made, this information can either be presented in a way that enhances decision-making or biases it. Therefore, better understanding how clinical feedback influences clinical decision-making will facilitate the design of clinical feedback systems (a.k.a. "decision support systems" in information processing literature) that aim to support clinicians in their practice (Browne, Pitts, & Wetherbe, 2007; Montgomery, Hosanager, Krishnan, & Clay, 2004; Payne, Bettman, & Johnson, 1993). The following are studies that propose future research to investigate the process of clinical decision-making and how feedback, with different presentations of information, influences this process.

Study A. The first proposed study includes a series of smaller studies that will explore the role of cognitive fit and better understand how the presentation of information may influence decision-making. Because this is a largely unexplored area of research in clinical psychology, exploratory work is necessary prior to conducting larger scale studies. These studies revolve around asking clinicians to share their thoughts, ideas, skills, and understandings of their clinical practice with researchers. As Kazdin (2008)

points out, the field of psychology can "profit enormously from codifying the experiences of clinicians in practice so that the information is accumulated and can be drawn on to generate and test hypotheses" (p. 115).

First, a series of cognitive interviews with practicing clinicians would be conducted that asks them to view different information presentations and to talk about how they interpret, understand, and would use that information. Clinicians would also be asked about whether they had a preference for one presentation over another and why. It is the sense of the author that the clinicians that completed the informal survey as part of this dissertation had thoughts and opinions about the survey content that they were not able to share within the survey responses. Engaging in dialogue about it would provide useful insight into this novel idea of categorical and dimensional presentations of information and whether tasks can be matched or mismatched with each presentation. One thing that would be useful to explore with clinicians, is the exact display that is used to present the information in feedback reports. For example, the hypothetical feedback report displaying dimensional information (Figure 23) included a horizontal line graph with an arrow showing where the youth's score was. This is only one of many ways to display this information; it could have also been displayed in a line graph or as a table with numbers. Understanding what display is most appealing, easiest to interpret, and/or fastest to incorporate into the clinician's mental representation would be extremely useful for designing future feedback reports.

Next, a group of experts would be gathered for a panel discussion about how useful hypothetical feedback reports with varying presentations of information are for making specific decisions about the client represented in the report. These experts would

represent an interdisciplinary peer-nominated panel of mental health professionals selected from across the nation. Each member would demonstrate their expertise by documenting extensive clinical or research credentials pertaining to their area of expertise. The panel would review hypothetical feedback reports that included the same information with different presentations and a consensus would be reached concerning what presentation they perceived was 'best' for informing each clinical task. Furthermore, the panel would review the information provided and determine what the specific decision or judgment would be for that client based on the feedback (e.g., is this client getting better? Would you refer this client for a medication consult? What level of care does this client need?). These answers can then be used to represent *best practice* and serve as a comparison to evaluate the accuracy of clinical decisions that will be assessed in the following study.

The last study in this group of exploratory studies utilizes the hypothetical cases reviewed by the expert panel and presents clinicians with a hypothetical feedback report for each client. The presentation of information on these feedback reports would differ by clinician, with some viewing categorical information, some dimensional, and some clinicians viewing both categorical and dimensional information. The clinicians would be asked to review each client's feedback report and make a determination for each decision or judgment. The clinician's answers would then be compared to the *best practice* decisions determined by the experts, allowing for the analysis of decision accuracy based on the presentation of information.

Another important feature of this final study would be to investigate the effect of time pressure on the accuracy of decision-making. Research demonstrates that accuracy

of decision-making is negatively affected by time pressure (e.g., Fraser-Mackenzie & Dror, 2011; Kerstholt, 1994; Speier et al., 2003). Clinicians often have large caseloads and may have little time outside of the therapy hour to write case notes and review feedback; they likely operate under some amount of time pressure. Therefore, it may be especially important to investigate how time pressure influences the cognitive fit of information presentation for decision-making. For example, with ample time, clinicians might make more accurate decisions based on feedback with both information presentations because it provides more detailed information that the clinician is able to select, interpret, and sort before making a decision. Yet, the inclusion of more information may also increase the cognitive load of the clinician, particularly when there is an influence of time pressure. Under those conditions, less information may be important to ensure the cognitive load is not too high. Thus, the goal of this proposed study would be to create a level of time pressure that represents what is typical during clinical practice. In other words, clinicians would be asked to make decisions quickly without having limitless amounts of time to review the feedback. Similar to Speier et al.'s (2003) hypothesis, there may even be a crossover effect where a particular presentation of information supports more accurate decisions when there is no time pressure, but harms accuracy once a level of time pressure is reached.

Study B. In addition to knowing very little about how the presentation of feedback affects decision-making, there is a similar lack of information about how clinical decisions are made in general (Falvey, et al., 2005; Spengler et al., 2009; Street et al., 2000), specifically the cognitive processes underlying them. Historically, studies investigating the cognitive processes underlying individual decision-making have used

one of two methodologies: structural or procedural approaches (Riedl, Bransdtätter, & Roithmayr, 2008). Structural approaches describe the relation between the information presented (i.e., input or stimuli) and the decision response (i.e., output) in order to make inferences about the decision-making process. Procedural approaches, on the other hand, attempt to directly capture the cognitive process that occurs between the input and the output. This is often called *process tracing* (Todd & Benbasat, 1987). Consistent with strong recommendations from researchers from multiple fields to utilize a multimethod approach to explore decision-making (e.g., Costa-Gomes, Crawford, & Broseta, 2001; Harte & Koele, 2001; Riedl et al., 2008), this study proposes to use both structural and process-tracing approaches.

The purpose of this study is to use a structural as well as a thinking aloud process-tracing strategy to identify if and how clinical information presented in feedback is applied to specific clinical tasks. The thinking aloud strategy can be carried out both retrospectively and concurrently with the task. In the current proposed research, the concurrent method is recommended as empirical evidence has long demonstrated that retrospective thinking aloud yields unreliable data due to memory distortion, interpretation, and an inability to recall facts (Russo, Johnson, & Stephens, 1989; Todd & Benbasat, 1987).

In this study, licensed clinicians would be asked to read a simulated case of youth receiving mental health treatment from one of the cases included in The Clinical Treatment Planning Simulation (CTPS; Falvey, 1994; Falvey, Bray, & Hebert, 2005). The cases in the CTPS were developed by a panel of mental health experts and include pieces of information about the case that are seen as important for treatment planning

such a intake interview, psychosocial history, academic record, medical history, client interview and parent interview summaries. After reviewing these materials, clinicians write a case conceptualization and treatment plan that are scored by a weighted scoring procedure that was also developed and evaluated by expert panelists. As a result, clinicians receive two scores, one for case conceptualization and one for treatment planning, reflecting their content knowledge of the symptomology and treatment practices for the specific client issue presented in the case. The CTPS cases have demonstrated content and construct validity in several studies (Falvey, 1992, 2001; Falvey et al., 2005). The goals of using the CTPS cases and scoring rubric are: 1) to provide comprehensive and standardized cases to clinicians that are consistent with actual clinical practice; and 2) to provide a baseline quantitative score reflecting the quality of the clinical decisions made.

After the CTPS case is reviewed and scored, The Structured Follow-up Interview developed by Falvey and colleagues (2005) would be used to better understand the influences on their decision process in case conceptualization and treatment planning. For example, the interview includes questions about what material provided in the case was most/least influential for treatment planning and asks the clinician to rank order treatment priorities and what specific information was used for determining them. Similar to the process described by Falvey et al. (2005), these interviews can be reviewed and coded based on what specific material and pieces of information was/was not used when determining the treatment plan. For example, one could explore whether information about client strengths is used more or less than information about symptoms present at home and/or school. Similarly, one could explore whether medical history is

used more or less than formal academic testing scores. Furthermore, one could explore what sources of input are used more or less than others (e.g., parent input vs. client input). These types of comparisons would allow for a better understanding of what types of information are attended to when a case is conceptualized and treatment plan made.

Finally, clinicians would be provided with a hypothetical clinical feedback report, created by a panel of experts, to reflect the client's status after receiving mental health treatment for a specified length of time (e.g., two months). Based on the newly provided information in the feedback report, clinicians would be asked to make a series of specific clinical decisions about the client. These clinical tasks would be similar to those presented in Chapter 1 and that were included in the informal clinical survey utilized in the current dissertation. While making each of these decisions, the clinicians would be asked to "think aloud" about their decision-making process. This process is meant to be unobtrusive and the only interaction by the researcher would be to provide a neutral prompt to ask the clinician to verbalize if there had been an extended period of silence (e.g., more than 10 seconds). Although this procedure can yield a large volume of unstructured data, the unobtrusive and concurrent nature of this process may better capture the clinician's decision-making process as it occurs naturally within practice (i.e., without researcher interference or limits on cognitive recall).

Study C. It is estimated that up to 24% of children receiving mental health treatment leave treatment with more severe symptoms than they exhibited when they started (Warren, Nelson, Mondragon, Baldwin, & Burlingame, 2010). Research suggests this is because clinicians struggle to recognize when a client is deteriorating (Hatfield, McCullough, Frantz, & Krieger, 2010; Hatfield & Ogles, 2006; Lambert & Shimokawa,

2011). Thus, the theory behind the effectiveness of feedback systems is that they alert clinicians to when symptom severity is significantly worsening (Anker et al., 2009; Hatfield et al., 2010; Shimokawa, Lambert & Smart, 2010; Sparks, Kisler, Adams, & Blumen, 2011). Then, the clinician is able to carefully reconsider the treatment plan and alter it to better fit the needs of the client. The purpose of this first proposed study would be to investigate whether the presentation of the information in the feedback (i.e., as categorical or dimensional) matters in terms of its influence on clinical outcomes. A review of the literature indicated that this question this has yet to be investigated.

This proposed study utilizes Contextualized Feedback Systems® (CFS™), a measurement feedback system (MFS) that provides ongoing information about the process and progress of mental health treatment (Bickman et al., 2011). The feedback is created based on a set of clinically relevant measures found in the PTPB (including the SFSS) that are completed by the youth, caregiver, and clinician during the last five minutes of a clinical session. This data is electronically scored and analyzed and a computerized feedback report is immediately available for review by the clinician. CFS™ is the only MFS with demonstrated effectiveness in youth mental health settings (Bickman, et al., 2011). Bickman et al. (2011) found that youth whose clinicians received ongoing feedback reports improved faster in terms of symptom severity compared to youth whose clinician did not receive feedback.

To investigate whether differences in clinical outcome occur based on whether information in feedback is presented as categorical or dimensional, youth clients would be randomly assigned to one of four conditions upon intake to mental health treatment: 1) categorical feedback; 2) dimensional feedback; 3) both categorical and dimensional

feedback; or 4) no feedback. Per the recommended measurement schedule included in the PTPB, the SFSS would be completed after every clinical session by the youth, caregiver, and clinician from the beginning of treatment to termination. This data would be entered into CFS™ and automatically analyzed using a method assuming a categorical latent variable structure (i.e., LCA or LTA) for condition 1 and a dimensional latent variable structure (i.e., GRM analysis or longitudinal GRM analysis) for condition 2, and both methods for condition 3. The results of these analyses would provide the information provided to clinicians in their weekly feedback reports, similar to the format shown in aim 3. Thus, clinicians treating a youth in condition 1 would receive feedback with a categorical presentation (i.e., Figure 22); clinicians treating a youth in condition 2 would receive feedback with a dimensional presentation (i.e., Figure 23); and clinicians treating a youth in condition 3 would received feedback with both presentations. Clinicians treating youth in condition 4 would receive no feedback, although SFSS data would be collected on the same schedule as the other three conditions.

Similar to methods used by Bickman and colleagues (2011), the proposed study would use hierarchical linear modeling to assess whether the clinical outcomes of the youth (the direction or slope of symptom change) varies based on experimental condition. Since research suggests that a key role of feedback is alerting the clinician to when a youth is deteriorating, this match between information presentation and clinical task about changes in symptom severity will allow for the most effective and efficient decision-making on the part of the clinician, something that is essential to the therapeutic process. Based on the theory of cognitive fit, it is hypothesized that youth in condition 2 (dimensional presentation in feedback) will show better outcomes compared to the other

conditions. This hypothesis is also consistent with clinician preference for a dimensional presentation for use in assessing client symptom change as discussed in aim 3. If clinicians naturally think with a dimensional perspective when assessing youths' change over time (Helzer, et al., 2006), then information presented dimensionally will display the best cognitive fit with the task of assessing whether a youth is improving or deteriorating. Although dimensional information would also be included in the condition 3, it is hypothesized that the combination of information presentations results in more information that the clinician must work to integrate into their mental representation. This additional information may increase the cognitive load, reducing the effectiveness and efficiency of decision-making.

Although the proposed study above would inform the question about whether the presentation of information in feedback affects clinical outcomes, it does not provide any information about *how* the information influences clinical decisions. In fact, very little is known about how feedback is used in clinical decision-making (Falvey, et al., 2005). Therefore, in connection to the proposed randomized trial, a random sample of clinicians from the three feedback conditions would be selected to complete an additional questionnaire that asks questions targeting how the information presented in the feedback was/was not used to make any clinical decisions and the usefulness of that presentation for specific clinical tasks. This questionnaire would be administered through CFS™ immediately after a clinician viewed a feedback report so that it was connected to a specific clinical session and report (i.e., while the information was fresh and clinical decisions actively being made). Alternatively, clinical supervisors could conduct structured interviews with clinicians during supervision meetings and discuss treatment

planning with the feedback report in front of them. This line of research would allow for the understanding of how feedback influences decision-making within the context in which clinical decisions are made, something that has been repeatedly stressed as important (e.g., Greenberg, 1991; Marmar, 1990; Schottenbauer, Glass, & Arnkoff, 2007; Shoham-Salomon, 1990).

General Conclusion and Final Thoughts

The discussion amongst researchers and practitioners about whether latent clinical variables are categorical or dimensional in nature is ongoing. Currently, evidence both for and against each approach exists. However, there is great similarity between these approaches, as is seen by the presence ordered latent classes along an underlying continuum supports. In fact, a mixed (or "hierarchical") approach has also been suggested and supported in the literature, where a variable is modeled both categorically and dimensionally (Bauer & McNaughton Reyes, 2007; Krueger et al., 2005; Markon & Krueger, 2005; Walton et al., 2011). Therefore, externalizing symptom severity may be conceptualized as having distinct latent classes (i.e., qualitative differences), each of which contains a dimensional aspect representing within-class heterogeneity (i.e., quantitative differences). Although model selection should be based on its reasonable representation of reality, the true structure of these latent variables are unknown. In fact, determining whether variation in clinical variables is truly categorical or dimensional (or both) is quite difficult and requires programmatic construct validation research (Bauer & Curran, 2003a; 2003b; Muthén, 2003).

Currently, when deciding what type of statistical model to use with clinical data, researchers often select one based on an a priori preference for whichever model matches with their theoretical model of the construct of the clinical variable. Thus, a model assuming a categorical latent variable would be selected if one believes that individuals differ qualitatively in terms of their externalizing severity, and a model assuming a dimensional latent variable would be selected if one believes individuals differ quantitatively. Application of the model would then be used as proof for the particular theoretical model, leading to confirmation bias (e.g., I think externalizing severity is made up of three distinct latent groups; my LCA results confirm three groups, externalizing severity must be categorical).

Another approach to model selection is by model comparison. Here multiple models are fit to the data and results are compared based purely on statistical fit. This approach has resulted in published research advocating one model over another based on model fit statistics (e.g., Krueger et al., 2005; Walton et al., 2011). However, comparing statistical fit does not answer the question about whether the individual differences in clinical variables are quantitative (i.e., categorical) or quantitative (i.e., dimensional) in nature. One reason for this is that it is difficult to separate assumptions made during modeling from the results used to draw conclusions about the theoretical structure of the construct. For example, if a measure displays significant floor or ceiling effects, thus violating an assumption of within-class normality, statistical fit indices may favor the presence of latent groups (i.e., categorical method), even when the variation in the data is purely quantitative (i.e., dimensional) in nature (Bauer, 2007). For reasons such as this,

Bauer and McNaughton Reyes (2007) point out, “it can be rather difficult to empirically adjudicate between the approaches on the basis of statistical fit alone” (p. 121).

Arriving at a definitive conclusion concerning the true structure of latent clinical variables is not likely to happen in the near future, if ever. In fact, incorporation of a dimensional perspective in the categorical DSM (i.e., revisions being made to the DSM-V to be released in May 2013) will likely to result in even more discussion about how best to conceptualized psychopathology and the implications these conceptualizations have for clinical practice. However, clinicians and researchers will continue to utilize clinical data to better understand and treat psychopathology. Therefore, it is important to understand how assumption made by the statistical models affects clinical decision-making. Even though the results of using a categorical or dimensional approach may provide similar overall messages to clinical concerning the symptom severity of youth, the presentation and interpretation of the information provided in feedback differs.

This dissertation laid the groundwork for adding a point of comparison (in addition to theoretical discussion and statistical comparison) when deciding whether to analyze clinical variables assuming a categorical or dimensional latent variable structure: the affect the latent variable structure has on clinical decision-making. The theory of cognitive fit suggests that the model that produces output for feedback in a format that promotes the most efficient and effective clinical decisions should be used. Thus, despite the similarity between approaches, the categorical presentation is more conducive to decisions made about group differences and the dimensional presentation is more conducive to decisions made about graded individual differences. In this way, clinical

decision-making can be optimized by matching the information presentation with the clinical task at hand.

Clinicians make important decisions throughout the mental health treatment process. Their judgments are central and client outcomes are reliant on a clinician's ability to make accurate decisions. Systematic evaluation and the use of feedback systems are meant to complement and support clinical judgment. As the body of research grows providing evidence for their effectiveness, it is possible that the use of treatment progress instruments will become a requirement for reimbursement by managed care companies in the future (Lutz et al., 2011; Stewart & Chambless, 2008). In this way, understanding how clinical feedback, specifically with different information presentations, affects clinical decision-making is important for ensuring that clinical practice is being supported, and not hindered, by the assumptions made during data analysis.

REFERENCES

Aboraya, A. (2007). The reliability of psychiatric diagnoses: POINT—our psychiatric diagnoses are still unreliable. *Psychiatry, 4*, 22 – 23. PMCID: PMC2922387

Achenbach, T. M. (1985). Assessment and taxonomy of child and adolescent psychopathology. Beverly Hills, CA: Sage.

Achenbach, T.M. (1991). *Manuel for the Youth Self-Report and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Edelbrock, C.S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85*, 1275-1301. doi: 10.1037/0033-2909.85.6.1275

Aggen, S.H., Neale, M.C., & Kendler, K.S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35*, 475 – 487. doi: 10.1017/S0033291704003563

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., Text Revision). Washington, DC: Author.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. doi: 10.1007/BF02294143

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. doi: 10.1007/BF02293814

Anker, M.G., Duncan, B.L., & Sparks, J.A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, 77(4), 693-704. doi: 10.1037/a0016062

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Athay, M.M. (2012). Satisfaction with Life Scale (SWLS) in caregivers of clinically-referred youth: Psychometric properties and mediation analysis. *Administration and Policy in Mental Health and Mental Health Services Research, 39*, 41 – 50. doi: 10.1007/s10488-011-0390-8

Athay, M.M., & Bickman, L. (2012). Development and psychometric evaluation of the youth and caregiver Service Satisfaction Scale. *Administration and Policy in*

Mental Health and Mental Health Services Research, 39, 71 – 77. doi: 10.1007/s10488-012-0407-y

Bauer, D.J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42,* 757 – 786. doi: 10.1080/00273170701710338

Bauer, D.J., & McNaughton Reyes, H.L. (2010). Modeling variability in individual development: Differences of degree or kind? *Child Development Perspectives,* 4, 114 – 122. doi: 10.1111/j.1750-8606.2010.00129.x

Bauer, D.J., & Curran, P.J. (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8,* 338 – 363. doi: 10.1037/1082-989X.8.3.338

Bauer, D.J., & Curran, P.J. (2003b). Overextraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthen (2003), and Cudeck and Henly (2003). *Psychological Methods, 8,* 384 – 393. doi: 10.1037/1082-989X.8.3.384

Bell, I., & Mellor, D. (2009). Clinical judgements: Research and practice. *Australian Psychologist, 44,* 112 – 121. doi: 10.1080/00050060802550023

Benbasat, I., & Taylor, R.N. (1982). Behavioral aspects of information processing or the design of management information systems. *IEEE Transactions on Systems Man and Cybernetics,* 12, 438 – 450. doi: 10.1109/TSMC.1982.4308848

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246. doi: 10.1037/0033-2909.107.2.238

Bettman, J.R., & Kakkar, P. (1977). Effects of information presentation format on consumer information acquisition strategies. *Journal of Consumer Research, 3,* 233 – 240. doi: 10.1086/208672

Bickman, L., Athay, M.M., Riemer, M., Lambert, E.W., Kelley, S.D., Breda, C., Tempesti, T., Dew-Reeves, S.E., Brannan, A.M., Vides de Andrade, A.R. (eds). (2010). *Manual of the Peabody Treatment Progress Battery, 2nd ed.* [Electronic version]. Nashville, TN: Vanderbilt University. http://peabody.vanderbilt.edu/ptpb\

Bickman, L., Kelley, S., Breda, C., Vides de Andrade, A.R., & Riemer, M. (2011). Effects of routine feedback to clinicians on youth mental health outcomes: A randomized cluster design. *Psychiatric Services, 62,* 1423 – 1429. doi: 10.1176/appi.ps.002052011

Bickman, L., Riemer, M., Lambert, E.W., Kelley, S.D., Breda, C., Dew, S.E., Brannan, A.M., Vides de Andrade, A.R. (eds). (2007). *Manual of the Peabody Treatment Progress Battery* [Electronic version]. Nashville, TN: Vanderbilt University. http://peabody.vanderbilt.edu/ptpb\

Bickman, L., Wighton, L.G., Lambert, E.W., Karver, M.C., & Steding, L. (2012). Problems in using diagnosis in child and adolescent mental health services research. *Journal of Methods and Measurement in the Social Sciences*, 3, 1 – 26. https://journals.uair.arizona.edu/index.php/jmmss

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*, 1089-1108. doi: 10.1002/jclp.20503

Brannan, A.M., Heflinger, C.A., & Bickman, L. (1997). The Caregiver Strain Questionnaire: Measuring the impact on the family of living with a child with serious emotional disturbance. *Journal of Emotional and Behavioral Disorders*, *5*, 212-222. doi: 10.1177/106342669700500404

Brown, T.A., & Barlow, D.H. (2005). Categorical vs dimensional classification of mental disorders in DSM-V and beyond. *Journal of Abnormal Psychology*, *114*, 551 – 556. doi: 10.1037/0021-843X.114.4.551

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K. A. & Long, J. S. (Eds.) *Testing Structural Equation Models*. pp. 136–162. Beverly Hills, CA: Sage.

Browne, G. J., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, 31, 89–104. http://misq.org/

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi: 10.1037/h0046016

Celeux, G., & Soromehno, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195 – 212. doi: 10.1007/BF01246098

Chandra, A., & Krovi, R. (1999). Representational congruence and information retrieval: Towards an extended model of cognitive fit. *Decision Support Systems*, *25*, 271 – 288. doi: 10.1016/S0167-9236(99)00014-7

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement*, *34*, 583 - 604. doi: 10.1177/0146621610362978

Chodoff, P. (2002). The medicalization of the human condition. *Psychiatric Services, 53,* 627 – 628. doi: 10.1176/appi.ps.53.5.627

Clogg, C. C. (1995). Latent class models. In G. Arminger, C.C. Clogg, & M.E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (Ch. 6: pp. 311 – 359). New York: Plenum.

Coghil, D., & Sonuga-Barke, E.J.S. (2012). Annual research review: Categories versus dimensions in the classification and conceptualization of child and adolescent mental disorders – implications of recent empirical study. *Journal of Child Psychology and Psychiatry, 53,* 469 – 489. doi:10.1111/j.1469-7610.2011.02511.x

Cole, D.A., Cai, L., Martin, N.C., Findling, R.L., Youngstrom, E.A., Garber, J., … Forehand, R. (2011). Structure and measurement of depression in youths: Applying item response theory to clinical data. *Psychological Assessment, 23,* 819 – 833. doi: 10.1037/a0023518

Collins, L. M., & Lanza, S. T., (2010). *Latent Class and Latent Transition Analysis for the Social, Behavioral, and Health Sciences*. New York: Wiley.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27,* 131-157. doi: 10.1207/s15327906mbr2701_8

Cooper, A., & Gomez, R. (2008). The development of a short form of the Sensitivity to Punishment and Sensitivity to Reward Questionnaire. *Journal of Individual Differences, 29,* 90 – 104. doi: 10.1027/1614-0001.29.2.90

Costa-Gomes, M., Crawford, V.P., & Bruno, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica, 69,* 1193 – 1235. doi: 10.1111/1468-0262.00239

Cuthbert, B.N. (2005). Dimensional models of psychopathology: Research agenda and clinical utility. *Journal of Abnormal Psychology¸ 114,* 565-569. doi: 10.1037/0021-843X.114.4.565

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science,* 243, 1668 – 1674. doi: 10.1126/science.2648573

De Clercq, B., De Fruyt, F., Van Leeuwen, K., & Mervielde, I. (2006). The structure of maladaptive personality traits in childhood: A step toward an integrative developmental perspective for DSM-V. *Journal of Abnormal Psychology, 115,* 639-657. doi: 10.1037/0021-843X.115.4.639

de Nijs, P.F.A., van Lier, P.A.C., Verhulst, F.C., & Ferdinand, R.F. (2007). Classes of disruptive behavior problems in referred adolescents. *Psychopathology, 40,* 440 – 445. doi: 10.1159/000107428

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39,* 1 – 38. http://www.jstor.org/

Dew-Reeves, S. E., & Athay, M.M. (2012). Validation and use of the youth and caregiver Treatment Outcomes Expectations Scale (TOES) to assess the relationship between expectations, pretreatment characteristics, and outcomes. *Administration and Policy in Mental Health and Mental Health Services Research, 39,* 90 – 103. doi: 10.1007/s10488-012-0406-z

Diener, E., Emmons, R.A., Larson, R.J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49,* 71-75. doi: 10.1207/s15327752jpa4901_13

Diener, E., Suh, E.M., Lucas, R.E., & Smith, H.L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125,* 276–302. doi: 10.1037/0033-2909.125.2.276

Dumenci, L., & Achenbach, T.M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment,* 20, 55 – 62. doi: 10.1037/1040-3590.20.1.55

Eaves, L.J., Silerg, J.L., Hewitt, J.K., Rutter, M., Meyer, J.M., Neale, M.C., et al. (1993). Analyzing twin resemblance in multisymptom data: Genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys. *Behavior Genetics, 23,* 5 – 19. doi: 10.1007/BF01067550

Eddy, K.T., Dorer, D.J., Franko, D.L., Tahilani, K., Thompson-Brenner, H., & Herzog, D.B. (2008). Diagnostic crossover in anorexia nervosa and bulimia nervosa: Implications for DSM-V. *The American Journal of Psychiatry, 165,* 245-250. doi: 10.1176/appi.ajp.2007.07060951

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology, 81,* 869–885. doi: 10.1037/0022-3514.81.5.869

Ekas, N., & Whitman, T.L. (2010). Autism symptom topography and maternal socioemotional functioning. *American Journal on Intellectual and Developmental Disabilities, 115,* 234-249. doi: 10.1352/1944-7558-115.3.234

Elia, J., Ambrosini, P., & Berrettini, W. (2008). ADHD characteristics: I. Concurrent co-morbidity patterns in children & adolescents. *Child and Adolescent Psychiatry and Mental Health*, 2. doi: 10.1186/1753-2000-2-15

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349. doi: 10.1037/1040-3590.8.4.341

Enders, C.K., & Bandalos, D.L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 430 – 457. doi: 10.1207/S15328007SEM0803_5

Falvey, J. E. (1992). From intake to intervention: Interdisciplinary perspectives on mental health treatment planning. *Journal of Mental Health Counseling*, 14, 471-489. http://www.amhca.org/news/journal.aspx

Falvey, J. E. (1994). [Clinical Treatment Planning Simulation: ADHD]. Unpublished document, University of New Hampshire. Clinical Judgment 34

Falvey, J. E. (2001). Clinical judgment in case conceptualization and treatment planning across mental health disciplines. *Journal of Counseling and Development, 79,* 292 - 303. doi: 10.1002/j.1556-6676.2001.tb01974.x

Falvey, J.E., Bray, T.E., & Herbert, D.J. (2005). Case conceptualization and treatment planning: Investigation of problem-solving and clinical judgment. *Journal of Mental Health Counseling*, 27, 348 – 372. http://www.amhca.org/news/journal.aspx

Falzer, P.R., & Garman, D.M. (2012). Image theory's counting rule in clinical decision making: Does it describe how clinicians make patient-specific forecasts? *Judgment and Decision Making, 7*, 268 – 281. http://journal.sjdm.org

Faust, D. (1984). *The limits of scientific reasoning*. As cited in Bell, I., & Mellor, D. (2009). Clinical judgements: Research and practice. *Australian Psychologist, 44*, 112 – 121. doi: 10.1080/00050060802550023

Fishler, E., Grossmann, M., & Messer, H. (2002). Detection of signals by information theoretic criteria: General asymptotic performance analysis. *IEEE Transactions on Signal Processing, 50,* 1027 – 1036. doi: 10.1109/78.995060

Fishman, D.B. (2001). From single case to database: A new method for enhancing psychotherapy, forensic, and other psychological practice. *Applied and Preventive Psychology*, 10, 275 – 304. doi: 10.1016/S0962-1849(01)80004-4

Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: The prevalence of DSM-IV disorders. *Journal of the*

American Academy of Child and Adolescent Psychiatry, 42, 1203 – 1211. doi: 10.1097/00004583-200310000-00011

Fraser-Mackenzie, P.A., & Dror, I.E. (2011). Dynamic reasoning and time pressure: Transition from analytical operations to experiential response. *Theory and Decision, 71,* 211 – 225. doi: 10.1007/s11238-009-9181-z

Gadermann, A.M., Schoner-Reichl, K.A. & Zumbo, B.D. (2010). Investigating validity evidence of the satisfaction with life scale adapted for children. *Social Indicators Research, 96,* 229–247. doi: 10.1007/s11205-009-9474-1

Gambrill, E. (2005). *Critical thinking in clinical practice: Improving the quality of judgments and decisions* (2nd ed.). Hoboken, NJ: John Wiley.

Garb, H.N. (2000). Computers will become increasingly important for psychological assessment: Not that there's anything wrong with that! *Psychological Assessment, 12,* 31 – 39. doi: 10.1037/1040-3590.12.1.31

Garb, H.N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1,* 67 – 89. doi: 10.1146/annurev.clinpsy.1.102803.143810

Gelhorn, H., Hartman, C., Sakai, J., Mikulich-Gilbertson, S., Stallings, M., Young, S., et al., (2009). An item response theory analysis of DSM-IV conduct disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 48,* 42 – 50. doi: 10.1097/CHI.0b013e31818b1c4e

Goldberg, D. (2000). Plato versus Aristotle: Categorical and dimensional models for common mental disorders. *Comprehensive Psychiatry, 41,* 8-13. doi: 10.1016/S0010-440X(00)80002-4

Goldstein, H. A. R. V. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215 – 231. doi: 10.1093/biomet/61.2.215

Grayson, J. (1987). Can categorical and dimensional views of psychiatric illness be distinguished? *British Journal of Psychiatry, 151,* 355 – 361. doi: 10.1192/bjp.151.3.355

Greenberg, L.S. (1991). Research on the process of change. *Psychotherapy Research, 1,* 3 – 16. doi: 10.1080/10503309112331334011

Grosse, S.D., Flores, A.L., Ouyang, L., Robbins, J.M., & Tilford, J.M. (2009). Impact of spina bifida on parental caregivers: Findings from a survey of Arkansas families.

Journal of Child and Family Studies, 18, 574-581. doi: 10.1007/s10826-009-9260-3

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19 – 30. doi: 10.1037/1040-3590.12.1.19

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23,* 56-62. doi: 10.1136/jnnp.23.1.56

Harlow, L.L. (2005). *The Essence of Multivariate Thinking.* Mahwah, New Jersey: Lawrence Earlbaum Associates, Inc.

Harmon, S.C., Lambert, M.J., Smart, D.M., Hawkins, E., Nielsen, S.L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research,* 17(4), 379-392. doi: 10.1080/10503300600702331

Harte, J.M., & Koele, P. (2001). Modeling and describing human judgment processes: The multiattribute evaluation case. *Thinking & Reasoning, 7,* 29 – 49. doi: 10.1080/13546780042000028

Hartman, C.A., Hox, J., Mellenbergh, G.J., Boyle, M.H., Offord, D.R., Racine, Y., et al. (2001). DSM-IV internal construct validity: When a taxonomy meets data. *Journal of Child Psychology and Psychiatry, 42,* 817 – 836. doi: 10.1111/1469-7610.00778

Hastings, R.P., Daley,D., Burns, C., Beck, A., & MacLean, W.E. Jr. (2006). Maternal distress and expressed emotion: Cross-sectional and longitudinal relationships with behavior problems of children with intellectual disabilities. *American Journal of Mental Retardation, 111,* 48-61. doi: 10.1352/0895-8017(2006)111[48:MDAEEC]2.0.CO;2

Hatfield, D., McCullough, L., Frantz, S.H.B., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology and Psychology*, 17, 25 – 32. doi: 10.1002/cpp.656

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35,* 485–491. doi: 10.1037/0735-7028.35.5.485

Heltzer, J.E., & Hudziak, J.J. (eds) (2002). *Defining psychopathology in the 21st century: DSM-V and Beyond.* As cited in Heltzer, J.E., Kraemer, H.C., & Krueger, R.F.

(2006). The feasibility and need for dimensional psychiatric diagnoses. *Psychological Medicine, 36,* 1671-1680. doi: 10.1017/S003329170600821X

Heltzer, J.E., Kraemer, H.C., & Krueger, R.F. (2006). The feasibility and need for dimensional psychiatric diagnoses. *Psychological Medicine, 36,* 1671-1680. doi: 10.1017/S003329170600821X

Hoagwood, K., & Kolko, D.L. (2009). Introduction to the special section on practice contexts: A glimpse into the nether world of public mental health services for children and families. *Administration and Policy in Mental Health and Mental Health Services, 36,* 35 – 36. doi: 10.1007/s10488-008-0201-z

Holt, R.R. (1986). Clinical and statistical prediction: A retrospective and would be integrative perspective. *Journal of Personality Assessment, 50,* 376 – 386. doi: 10.1207/s15327752jpa5003_7

Hu & Bentler (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi: 10.1080/10705519909540118

Hubley, A.M., & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123,* 207–215. doi: 10.1080/00221309.1996.9921273

Hwang, M.I. (1994). Decision making under time pressure: A model for information systems research. *Information and Management, 27,* 197 – 203. doi: 10.1016/0378-7206(94)90048-5

Jones, K.D. (2012). Dimensional and cross-cutting assessment in the DSM-5. *Journal of Counseling and Development, 90,* 481 – 487. doi:10.1002/j.1556-6676.2012.00059.x

Joreskog, K.G. (1988). Analysis of covariance structures. In J.R. Nesselroade and R.B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 207-230). New York: NY, Plenum Press. doi: 10.1007/978-1-4613-0893-5_5

Judd, L.L., Akiskal, H.S., & Paulus, M.P. (1997). The role and clinical significance of subthreshold depressive symptoms (SSD) in unipolar major depressive disorder. *Journal of Affective Disorders, 45,* 5-17. doi: 10.1016/S0165-0327(97)00055-4

Judd, L.L., Shettler, P.J., Akiskal, H.S., Coryell, W., Leon, A.C., Maser, J.D. et al. (2008). Residual symptom recovery from major affective episodes in bipolar disorders and rapid episode relapse/recurrence. *Archives of General Psychiatry, 65,* 386-394. doi: 10.1001/archpsyc.65.4.386

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151. doi: 10.1177/001316446002000116

Kamphuis, J.H., & Noordhof, A. (2009). On categorical diagnoses in DSM-V: Cutting dimensions at useful points? *Psychological Assessment, 21,* 294-301. doi: 10.1037/a0016697

Kay, S.R. Fisbein, A., & Opler, L.A. (1987). The Positive and negative syndrome scale (PANASS) for schizophrenia. *Schizophrenia Bulletin, 13,* 261-276. doi: 10.1093/schbul/13.2.261

Kazdin, A.E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance knowledge base, and improve patient care. *American Psychologist,* 63, 146 – 159. doi: 10.1037/0003-066X.63.3.146

Kerstholt, J.H. (1994). The effect of time pressure on decision-making behavior in a dynamic task environment. *Acta Psychologica, 86,* 89 – 104. doi: 10.1016/0001-6918(94)90013-2

Kessler, R.C., Demler, O., Frank, R.G., Olfson, M., Pincus, H.A., Walters, E.E. et al. (2005). Prevalence and treatment of mental disorders, 1990 to 2003. *The New England Journal of Medicine, 16,* 2515-2523. doi: 10.1056/NEJMsa043266

Kim, S., Kim, S-H., & Kamphaus, R.W. (2010). Is aggression the same for boys and girls? Assessing measurement invariance with confirmatory factor analysis and item response theory. *School Psychology Quarterly, 25,* 45 – 61. doi: 10.1037/a0018768

Klein, D.N., & Riso, L.P. (1993). Psychiatric disorders: Problems of boundaries and comorbidity. In C.G. Costello (Ed.), *Basic issues in psychopathology* (pp. 16 – 66). New York: Guliford Press.

Kliger, D., & Kudryavtsev, A. (2010). The availability heuristic and investors' reaction to company-specific event. *Journal of Behavioral Finance,* 11, 50 – 65. doi: 10.1080/15427561003591116

Kluger, A.N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin,* 119, 254 – 284. doi: 10.1037/0033-2909.119.2.254

Kooijmans, R., Scheres, A., & Oosterlaan, J. (2000). Response inhibition and measures of psychopathology: A dimensional analysis. *Child Neuropsychology, 6,* 175-184. doi: 10.1076/chin.6.3.175.3154

Krueger, R.F., Markon, K.E., Patrick, C.J., & Iacono, W.G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology, 114,* 537-550. doi: 10.1037/0021-843X.114.4.537

Krueger, R.F., Watson, D., & Barlow, D.H. (2005). Introduction to the special section: Toward a dimensionally based taxonomy of psychopathology. *Journal of Abnormal Psychology, 114,* 491-493. doi: 10.1037/0021-843X.114.4.491

Lacourse, E., Baillargeon, R., Dupere, V., Vitaro, F., Romano, E., & Tremblay, R. (2010). Two-year predictive validity of conduct disorder subtypes in early adolescence: A latent class analysis of a Canadian longitudinal sample. *Journal of Child Psychology and Psychiatry, 51,* 1386 – 1394. doi: 10.1111/j.1469-7610.2010.02291.x

Lambert, C.M., Essau, C.A., Schmitt, N., & Samms-Vaughan, M.E. (2007). Dimensionality and psychometric invariance of the Youth Self-Report form of the Child Behavior Checklist in cross-national settings. *Assessment, 14,* 231 – 245. doi: 10.1177/1073191107302036

Lambert, M.J., Gregersen, A.T., & Burlingame, G.M. (2004). The Outcome Questionnaire-45. In Lambert, Gregersen & Burlingame (Ed.), *The Use of Psychological Testing for Treatment Planning and Outcome Assessment: Volume 3: Instruments for Adults (3rd ed;* pp. 191 – 234). Mahwah, NJ: Lawrence Erlbaum Associates.

Lambert, M.J., Harmon, C., Slade, K., Whipple, J.L., & Hawkins, E.J. (2005). Providing feedback to psychotherapists on their patients' progress: clinical results and practice suggestions. *Journal of Clinical Psychology,* 61, 165-174. doi: 10.1002/jclp.20113

Lambert, M.J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48,* 72 – 79. doi: 10.1037/a0022238

Lambert, M.J., Whipple, J.L., Vermeersch, D.A., Smart, D.A., Hawkins, E.J., et al. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9,* 91 – 103. doi: 10.1002/cpp.324

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent Structure Analysis,* Boston: Houghton Mifflin.

Leve, L.D., Kim, H.K., & Pears, K.C. (2005). Childhood temperament and family environment as predictors of internalizing and externalizing trajectories from ages

5 to 17. *Journal of Abnormal Child Psychology, 33*, 505-520. doi: 10.1007/s10802-005-6734-7

Little, R.J.A., & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons.

Lo, Y., Mendell, N.R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767 – 778. doi: 10.1093/biomet/88.3.767

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lutz, W., Böhnke, J.R., & Köck, K. (2011). Lending an ear to feedback systems: Evaluation of recovery and non-response in psychotherapy in a German outpatient setting. *Community Mental Health Journal*, 47, 311 – 317. doi: 10.1007/s10597-010-9307-3

Lutz, W., Lambert, M.J., Harmon, S.C., Tachitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure, and duration – What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy*, 13, 223 – 232. doi: 10.1002/cpp.496

Mack, A.H., Forman, L., Breown, R., & Francis, A. (1994). A brief history of psychiatric classification: From the ancients to DSM-IV. *Psychiatric Clinics of North America, 17*, 515 – 523.

Magidson, J., & Vermunt, J. K. (2000). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Markon, K.E., & Krueger, R.F. (2005). Categorical and continuous models of liability to externalizing disorders: A direct comparison in NESARC. *Archives of General Psychiatry, 62*, 1352 – 1359. doi: 10.1001/archpsyc.62.12.1352

Marmar, C.R. (1990). Psychotherapy process research: Progress, dilemmas, and future directions. *Journal of Consulting and Clinical Psychology, 58*, 265 – 272. doi: 10.1037/0022-006X.58.3.265

Maser, J.D., Norman, S.B., Zisook, S., Everall, I.P., Stein, M.B., Schettler, P.J., & Judd, L.L. (2009). Psychiatric nosology is ready for a paradigm shift in DSM-V. *Clinical Psychology: Science and Practice, 16*, 24-40. doi: 10.1111/j.1468-2850.2009.01140.x

Maughan, B. (2005). Developmental trajectory modeling: A view from developmental psychopathology. *The Annals of the American Academy of Political and Social Science, 602*, 119 – 130. doi: 10.1177/0002716205281067

McBride, O., & Adamson, G. (2010). Are subthreshold alcohol dependence symptoms a risk factor for developing DSM-IV alcohol use disorders? A three-year prospective study of 'diagnostic orphans' in a national sample. *Addictive Behaviors, 35,* 586-592. doi: 10.1016/j.addbeh.2010.01.014

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park: Sage Publications.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley.

McKnight, P.E., McKnight, K.M., Sidani, S., & Figueredo, A.J. (2007). *Missing Data: A Gentle Introduction*. New York, NY: Guilford Press.

Meehl, P.E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press.

Miller, T.A., & Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30, 107-122. doi: 10.1111/j.1745-3984.1993.tb01069.x

Millsap, R.E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4,* 5 – 9. doi: 10.1111/j.1750-8606.2009.00109.x

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334. doi: 10.1177/014662169301700401

Montgomery, A. L., Hosanagar, K., Krishnan, R., & Clay, K. B. (2004). Designing a better shopbot. *Management Science*, 50, 189–206. doi: 10.1287/mnsc.1030.0151

Morrow-Bradley, C. and Elliott, R. (1986). Utilization of psychotherapy research by practicing psychotherapists. *American Psychologist, 41,* 188-197. doi: 10.1037/0003-066X.41.2.188

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. Applied Psychological Measurement, 19, 73-90. doi: 10.1177/014662169501900109

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods, 8,* 369 – 377, 384 – 393. doi: 10.1037/1082-989X.8.3.369

Muthén, L.K. and Muthén, B.O. (2010). Mplus User's Guide. Sixth Edition.

Los Angeles, CA: Muthén & Muthén

National Institute of Mental Health (1999). *Bridging science and service: A report by the National Advisory Mental Health Council's Clinical Treatment and Services Research Workgroup*. Bethesda, MD: Author.

Narrow, W.E., & Kuhl, E.A. (2011). Dimensional approaches to psychiatric diagnosis in DSM-5. *Journal of Mental Health Policy and Economics, 14*, 197 – 2000. www.icmpe.org/test1/journal/journal.htm

Nock, M.K., Kazdin, A.E., Hiripi, E., & Kessler, R.C. (2006). Prevalence, subtypes, and correlates of DSM-IV conduct disorder in the National Comorbidity Survey Replication. *Psychological Medicine, 36*, 699 – 710. doi: 10.1017/S0033291706007082

Neto, F. (1993). The Satisfaction with Life Scale: Psychometric properties in an adolescent sample. *Journal of Youth and Adolescence, 22*, 125–134. doi: 10.1007/BF01536648

Normand, S-L., T., Belanger, A.J., & Eisen, S.V. (2006). Graded response model-based item selection for behavior and symptom identification. *Health Services and Outcome Research Methodology, 6*, 1 – 19. doi: 10.1007/s10742-006-0005-0

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1 – 18. doi: 10.1016/0022-2496(66)90002-2

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nylund, K. (2007). Latent transition analysis: Modeling extensions and an application to peer victimization. Doctoral dissertation, University of California, Los Angeles.

Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535 – 569. doi: 10.1080/10705510701575396

Odgers, C.L., Moretti, M.M., Burnette, M.L., Chauhan, P., Waite, D., & Reppucci, N.D. (2007). A latent variable modeling approach to indentifying subtypes of serious and violent female juvenile offenders. *Aggressive Behavior, 33*, 339 – 352. doi: 10.1002/ab.20190

Okasha, A. (2009). Would the use of dimensions instead of categories remove problems related to subthreshold disorders? *European Archives of Psychiatry and Clinical Neuroscience, 259*, S129-S133. doi: 10.1007/s00406-009-0052-y

Ostrander, R., Herman, K., Sikorski, J., Mascendaro, P., & Lambert, S. (2008). Patterns of psychopathology in children with ADHD: A latent profile analysis. *Journal of Clinical Child and Adolescent Psychology, 37,* 833 – 847. doi: 10.1080/15374410802359668

Pastor, D.A., & Beretvas, S.B. (2006). Longitudinal rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement, 30,* 100 – 120. doi: 10.1177/0146621605279761

Pavot, W., & Diener, E. (2008). The Satisfaction with Life Scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology, 3,* 137–152. doi: 10.1080/17439760701756946

Payne, J.W. (1976). Task complexity and contingent processing in decision making: A replication and extension. *Organizational Behavior and Human Performance, 16,* 366 – 387. doi: 10.1016/0030-5073(76)90022-2

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Puschner, B., Steffon, S., Slade, M., Kaliniecka, H., Maj, M., Fiorillo, A., Munk-Jørgensen, P., et al., (2010). Clinical decision making and outcome in routine care for people with severe mental illness (CEDAR): Study protocol. *BMC Psychiatry, 10*(90). doi: 10.1186/1471-244X-10-90

Rapport, M.D., LaFond, S.V., & Sivo, S.A. (2009). Unidimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A rasch analysis. *Journal of Psychopathological Behavior Assessment, 31,* 309-319. doi: 10.1007/s10862-008-9125-x

Reese, R.J., Norsworthy, L.A., & Rowlands, S.R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy Theory, Research, Practice, Training* 46(4), 418-431. doi: 10.1037/a0017901

Reise, S.P., & Waller, N.G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5,* 27-48. doi: 10.1146/annurev.clinpsy.032408.153553

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Ridley, C.R., & Shaw-Ridley, M. (2009). Clinical judgment accuracy: From meta-analysis to metatheory. *The Counseling Psychologist,* 37, 400 – 409. doi: 10.1177/0011000008330830

Ridley, C.R., Tracy, M.L., Pruitt-Stephens, L., Wimsatt, M.K., & Beard, J. (2008). Multicultural assessment validity: The preeminent ethical issue in psychological assessment. In L.A. Suzuki & J.G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed., pp. 22 – 23). San Francisco: Jossey-Bass.

Riedl, R., Brandstätter, E., & Roithmayr, F. (2008). Identifying decision strategies: A process- and outcome-based classification method. *Behavior Research Methods, 40*, 795 – 807. doi: 10.3758/BRM.40.3.795

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory and Cognition, 17*, 759 – 769. doi: 10.3758/BF03202637

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85 – 100). New York: Singer.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178 – 200. doi: 10.1037/h0023624

Schmitt, J.S. & Di Fabio, R.P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology, 57*, 1008–1018. doi: 10.1016/j.jclinepi.2004.02.007

Schottenbauer, M.A., Glass, C.R., & Arnkoff, D.B. (2007). Decision making and psychotherapy integration: Theoretical considerations, preliminary data, and implications for future research. *Journal of Psychotherapy Integration, 17*, 225 – 250. doi: 10.1037/1053-0479.17.3.225

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461 – 464. doi: 10.1214/aos/1176344136

Shankman, S.A., Lewinsohn, P.M., Klein, D.N., Small, J.W., Seeley, J.R., & Altman, S.E. (2009). Subthreshold conditions as precursors for full syndrome disorders: A 15-year longitudinal study of multiple diagnostic classes. *Journal of Child Psychology and Psychiatry, 50*, 1485-1494. doi: 10.1111/j.1469-7610.2009.02117.x

Shimokawa, K., Lambert, M.J., & Smart, D.W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a

psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*, 298 – 311. doi: 10.1037/a0019247

Shoham-Salomon, V. (1990). Interrelating research processes of process research. *Journal of Consulting and Clinical Psychology, 58*, 217 – 225. doi: 10.1037/0022-006X.58.3.295

Shulte, D. (1997). Dimensions of outcome measurement. In H.H. Strupp, L.M. Horowitz & M.J. Lambert (Eds), *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery* (pp. 57 – 80). Washington, DC: American Psychological Association Press.

Slade, K., Lambert, M.J., Harmon, S.C., Smart, D.W., & Bailey, R. (2008). Improving psychotherapy outcome: the use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology and Psychotherapy*, 15, 287-303. doi: 10.1002/cpp.594

Snyder, D.K. (2000). Computer-assisted judgment: Defining strengths and liabilities. *Psychological Assessment, 12*, 52 – 60. doi: 10.1037/1040-3590.12.1.52

Sparks, J.A., Kisler, T.S., Adams, J.F., & Blumen, D.G. (2011). Teaching accountability: Using client feedback to train effective family therapists. *Journal of Marital and Family Therapy, 37*, 452 – 467. doi: 10.1111/j.1752-0606.2011.00224.x

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72 – 101. doi: 10.2307/1412159

Spengler, P.M., White, M.J., Ægisdóttir, S., & Maugherman, A.S. (2009). Time keeps on ticking: The experience of clinical judgment. *The Counseling Psychologist*, 37, 416 – 423. doi: 10.1177/0011000009332008

Speier, C. (2006). The influence of information presentation formats on complex task decision-making performance. *International Journal of Human Computer Studies, 64*, 1115 – 1131. doi: 10.1016/j.ijhcs.2006.06.007

Speier, C., Vessey, I., & Valacich, J. (2003). The effects of interruptions and information presentation formats on decision performance. *Decision Sciences, 34*, 771 – 797. doi: 10.1111/j.1540-5414.2003.02292.x

Steiger, J.H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduck and Glaser. *Structural Equation Modeling: A Multidisciplinary Journal, 7,* 149-162. doi: 10.1207/S15328007SEM0702_1

Stein, B.D., Kogan, J.N., Hutchison, S.L., Magee, E.A., & Sorbero, M.J. (2010). Use of outcomes information in child mental health treatment: Results from a pilot study. *Psychiatric Services, 61,* 1211 – 1216. doi: 10.1176/appi.ps.61.12.1211

Steward, R.E., & Chambless, D.L. (2007). Does psychotherapy research inform treatment decisions in private practice? *Journal of Clinical Psychology, 63,* 267 – 281. doi: 10.1002/jclp.20347

Stewart, R.E., & Chambless, D.L. (2008). Treatment failures in private practice: How do psychologists proceed? *Professional Psychology: Research and Practice, 39,* 176 – 181. doi: 10.1037/0735-7028.39.2.176

Stewart, R.E., & Chambless, D.L. (2010). Interesting practitioners in training in empirically supported treatments: Research reviews versus case studies. *Journal of Clinical Psychology, 66,* 73 – 95. doi: 10.1002/jclp.20630

Storr, C.L., Accornero, V.H., & Crum, R.M. (2007). Profiles of current disruptive behavior: Association with recent drug consumption among adolescents. *Addictive Behaviors, 32,* 248 – 264. doi: 10.1016/j.addbeh.2006.03.045

Street, L.L., Niederehe, G., & Lebowitz, B.D. (2000). Toward greater public health relevance for psychotherapeutic intervention research: An NIMH workshop report. *Clinical Psychology: Science and Practice,* 7, 127 – 137. doi: 10.1093/clipsy.7.2.127

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distribution.* Wiley: Chichester, New York.

Todd, P., & Benbasat, I. (1987). Process tracing methods in decision support systems research: Exploring the black box. *MIS Quarterly, 11,* 493 – 512. doi: 10.2307/248979

Van den Akker, A.L., Dekovic, M., Asscher, J.J., Shiner, R.C., & Prinzie, P. (2012). Personality types in childhood: Relations to latent trajectory classes of problem behavior and overreactive parenting across the transition into adolescence. *Journal of Personality and Social Psychology.* Advanced online publication. doi: 10.1037/a0031184

Vassar, M., Ridge, J.W., & Hill, A.D. (2008). Inducing score reliability from previous reports: An examination of life satisfaction studies. *Social Indic Research, 87,* 27–45. doi: 10.1007/s11205-007-9157-8

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences, 22,* 219 – 240. doi: 10.1111/j.1540-5915.1991.tb00344.x

Vessey, I. (1994). The effect of information presentation on decision making: A cost-benefit analysis. *Information and Management, 27*, 103 – 119. doi: 10.1016/0378-7206(94)90010-8

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research, 2,* 63 – 84. doi: 10.1287/isre.2.1.63

Vitterso, J., Biswas-Diener, R., & Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the satisfaction with life scale in Greenland and Norway. *Social Indicators Research, 74,* 327–348. doi: 10.1007/s11205-004-4644-7

Walton, K.E., Ormel, J., & Krueger, R.F. (2011). The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models. *Journal of Abnormal Psychology*, 39, 553 – 561. doi: 10.1007/s10802-010-9478-y

Warren, J.S., Nelson, P.L., Mondragon, S.A., Baldwin, S.A., & Burlingame, G.M. (2010). Youth psychotherapy change trajectories & outcome in usual care: Community mental health versus managed care. *Journal of Clinical and Consulting Psychology, 78*, 144 – 155. doi: 10.1037/a0018544

Widiger, T.A., & Samuel, D.B. (2005). Diagnostic categories or dimensions? A question for the diagnostic and statistical manual of mental disorders – fifth edition. *Journal of Abnormal Psychology*, *114*, 494-504. doi: 10.1037/0021-843X.114.4.494

Willis, C.E., & Holmes-Rovner, M. (2006). Integrating decision making and mental health interventions research: Research directions. *Clinical Psychology: Science and Practice, 13,* 9 – 25. doi: 10.1111/j.1468-2850.2006.00002.x

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence-Erlbaum, page 146.

Witkiewitz, K., King, K., McMahon, R.J., Wu, J., Luk, J., et al. (2013). Evidence for a multi-dimensional latent structural model of externalizing disorders. *Journal of Abnormal Child Psychology, 41*, 223 – 237. doi: 10.1007/s10802-012-9674-z

Woods, S.B., Farineau, H.M., & McWey, L.M. (2013). Physical health, mental health, and behavior problems among early adolescents in foster care. *Child: Care, Health and Development, 39*, 220 – 227. doi: 10.1111/j.1365-2214.2011.01357.x

Worthen, V., & Lambert, M. (2007). Outcome oriented supervision: Advantages of adding systematic client tracking to supportive consultations. *Counseling and Psychotherapy Research*, 7(1), 48-53. doi: 10.1080/14733140601140873

Wright, B.D. (1999). Fundamental measurement of psychology. In S.E. Embretson & S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65 – 104). Mahwah, NJ: Earlbaum.

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370. Retrieved March 10, 2011 from www.rasch.org/rmt/rmt83b.htm

Wu, C.F.J.(1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11,* 95–103. http://imstat.org/aos/

Wu, M.L. (2007) *ACER ConQuest version 2.0: Generalised item response modelling software / Margaret L. Wu ... [et al.]* ACER Press, Camberwell, Vic.

Yang, C. (1998). Finite mixture model selection with psychometric applications (Doctoral dissertation, University of California, Los Angeles, 1996). *Dissertation Abstracts International, 59,* 3421B.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement, 30,* 233-251. doi: 10.1111/j.1745-3984.1993.tb00425.x

APPENDIX A

Informal Clinician Survey

Information from clinical outcome measures can be presented in different ways, depending on how the data are analyzed. For example, a client's symptom severity can be presented categorically, where he/she is classified as having 'Clinical', 'Sub-Clinical', or 'Non-Clinical' levels of severity (Figure 1). Here, change is evident when a client moves from one category to another. Alternatively, a client's symptom severity can be presented dimensionally, where the client falls along a continuum that ranges from the lowest possible severity to the highest (Figure 2) and utilizes percentiles based on a larger sample to help interpret the location of the score (i.e. low, medium, high). Change is indicated by the direction of the movement of the score, and whether the amount of change is significant.

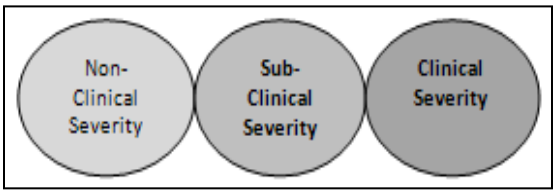


Figure 1. Categorical presentation

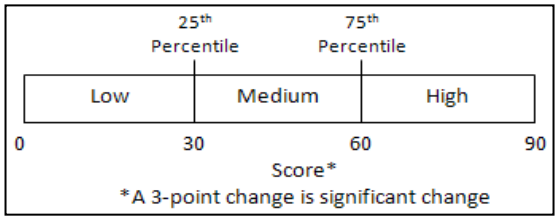
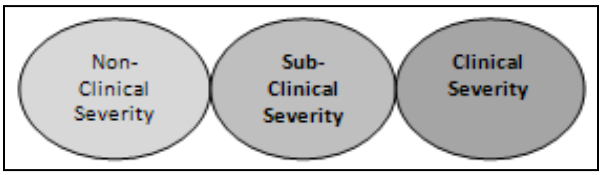


Figure 2. Dimensional presentation

The purpose of this informal survey is to explore what type of information presentation is most useful and easily applied to some specific clinical judgments and decisions made within treatment planning.

Based on published literature on clinical judgment and decision-making, several common tasks or questions that clinicians may address within the process of psychotherapy are listed below. For each one, please indicate how useful the given presentation format is for aiding in the general decision-making process of each task.

Categorical Presentation: Below is information about the client's symptom severity presented categorically. Every time the client completes the symptom severity measure, you (the clinician) receive an indication of which category the client is in.



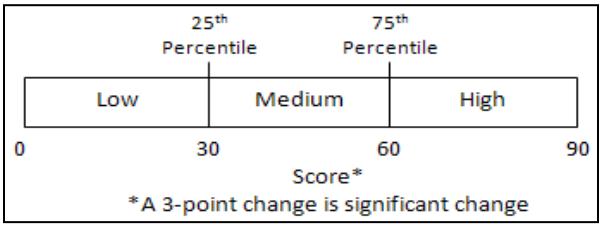
The table below depicts the information you might receive for four hypothetical clients who completed a measure of symptom severity at two time points.

Client	Time 1	Time 2
Sarah	Clinical	Sub-Clinical
Mary	Sub-Clinical	Sub-Clinical
Jane	Clinical	Non-Clinical
Anne	Sub-Clinical	Clinical

For each general decision/judgment listed, please indicate how *useful* this information would be for aiding in making clinical judgments and decisions.

How **useful** is the type of Information for aiding in determining…		Not Useful	A little Useful	Neutral	Useful	Very Useful
1.	… Does the client display clinical levels of psychopathology?	☐	☐	☐	☐	☐
2.	…What level of care is needed for the client?	☐	☐	☐	☐	☐
3.	…Should I assess or reassess the client for a diagnosis?	☐	☐	☐	☐	☐
4.	…Should the client be referred for medication consultation?	☐	☐	☐	☐	☐
5.	…Is the treatment working for the client?	☐	☐	☐	☐	☐
6.	…Is the client getting better?	☐	☐	☐	☐	☐
7.	…Should the client's treatment plan be changed?	☐	☐	☐	☐	☐
8.	…Is the client ready for termination of treatment?	☐	☐	☐	☐	☐
9.	…Is the client deteriorating?	☐	☐	☐	☐	☐
10.	…How frequent does this client need to have sessions?	☐	☐	☐	☐	☐
11.	…Does the client require hospitalization?	☐	☐	☐	☐	☐
12.	…Is this client developing new symptoms?	☐	☐	☐	☐	☐

Dimensional Presentation: Below is information about the client's symptom severity presented dimensionally. Every time the client completes the symptom severity measure, you (the clinician) receive a score reflecting his/her symptom severity.



The table below depicts the information you might receive for four hypothetical clients who completed a measure of symptom severity at two time points.

Client	Time 1 Score	Time 2 Score
Jill	82	75 (Improved)
Missy	58	56 (No Change)
Sally	62	31 (Improved)
Anna	59	72 (Deteriorated)

For each general decision/judgment listed, please indicate how *useful* this information would be for aiding in clinical judgments and decisions.

How **useful** is the type of Information for aiding in determining…		Not Useful	A little Useful	Neutral	Useful	Very Useful
1.	… Does the client display clinical levels of psychopathology?	☐	☐	☐	☐	☐
2.	…What level of care is needed for the client?	☐	☐	☐	☐	☐
3.	…Should I assess or reassess the client for a diagnosis?	☐	☐	☐	☐	☐
4.	…Should the client be referred for medication consultation?	☐	☐	☐	☐	☐
5.	…Is the treatment working for the client?	☐	☐	☐	☐	☐
6.	…Is the client getting better?	☐	☐	☐	☐	☐
7.	…Should the client's treatment plan be changed?	☐	☐	☐	☐	☐
8.	…Is the client ready for termination of treatment?	☐	☐	☐	☐	☐
9.	…Is the client deteriorating?	☐	☐	☐	☐	☐
10.	…How frequent does this client need to have sessions?	☐	☐	☐	☐	☐
11.	…Does the client require hospitalization?	☐	☐	☐	☐	☐
12.	…Is this client developing new symptoms?	☐	☐	☐	☐	☐

APPENDIX B

Symptoms and Functioning Severity Scale (SFSS-33)

This Youth's Behaviors, Thoughts, and Feelings (2020)

Below is a list of behaviors, thoughts, and feelings that youths may experience. Please put an 'X' in the one box that best matches how often you think this youth has experienced each of these things OVER THE LAST 2 WEEKS – either Never, Hardly Ever, Sometimes, Often or Very Often. When answering, think about the different places this youth may have experienced these things, for example, at school, at home, with friends, or at work (for older teens).

	IN THE LAST TWO WEEKS, HOW OFTEN DID THIS YOUTH:	Never	Hardly Ever	Some times	Often	Very Often
1.	. . . throw things when he/she was mad?	☐	☐	☐	☐	☐
2.	. . . eat a lot more or a lot less than usual?	☐	☐	☐	☐	☐
3.	. . . feel unhappy or sad?	☐	☐	☐	☐	☐
4.	. . . get into trouble?	☐	☐	☐	☐	☐
5.	. . . have little or no energy?	☐	☐	☐	☐	☐
6.	. . . disobey adults? (not do what adults told him/her to do)	☐	☐	☐	☐	☐
7.	. . . interrupt others?	☐	☐	☐	☐	☐
8.	. . . lie to get things he/she wanted?	☐	☐	☐	☐	☐
9.	. . . have a hard time controlling his/her temper?	☐	☐	☐	☐	☐
10.	. . . use drugs for non-medical purposes?	☐	☐	☐	☐	☐
11.	. . . worry about a lot of things?	☐	☐	☐	☐	☐
12.	. . . have a hard time getting along with family and/or friends?	☐	☐	☐	☐	☐

	IN THE LAST TWO WEEKS, HOW OFTEN DID THIS YOUTH:	Never	Hardly Ever	Some times	Often	Very Often
13.	. . . threaten or bully others?	☐	☐	☐	☐	☐
14.	. . . feel worthless?	☐	☐	☐	☐	☐
15.	. . . drink alcohol (beer, wine, hard liquor)?	☐	☐	☐	☐	☐
16.	. . . have a hard time having fun?	☐	☐	☐	☐	☐
17.	. . . feel afraid that other kids would laugh at him/her?	☐	☐	☐	☐	☐
18.	. . . have a hard time waiting his/her turn?	☐	☐	☐	☐	☐
19.	. . . sleep a lot more than he/she normally does?	☐	☐	☐	☐	☐
20.	. . . hang out with kids who get into trouble?	☐	☐	☐	☐	☐
21.	. . . feel nervous and/or shy around other people?	☐	☐	☐	☐	☐
22.	. . . have a hard time paying attention?	☐	☐	☐	☐	☐
23.	. . . get into fights with family members and/or friends?	☐	☐	☐	☐	☐
24.	. . . lose things he/she needs?	☐	☐	☐	☐	☐
25.	. . . have a hard time sitting still?	☐	☐	☐	☐	☐
26.	. . . have a hard time sleeping because he/she was worrying?	☐	☐	☐	☐	☐
27.	. . . feel tense?	☐	☐	☐	☐	☐
28.	. . . cry easily?	☐	☐	☐	☐	☐
29.	. . . annoy other people on purpose?	☐	☐	☐	☐	☐
30.	. . . argue with adults?	☐	☐	☐	☐	☐
31.	. . . think that he/she doesn't have any friends?	☐	☐	☐	☐	☐
32.	. . . feel too scared to ask questions in class? *If this youth has not been in class in the last two weeks, please answer how you think the youth might have felt if he/she were in school in the last two weeks.	☐	☐	☐	☐	☐

APPENDIX C

Annotated Mplus Syntax Used for Analyses

Latent Class Analysis Model

```
!Name of Data file
DATA:  FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;

!All the variables in the data file
VARIABLE:   NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
          T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
          T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
          T2_B12 T2_A15 T2_B15;
!variables used in analysis
         USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                  T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15;
!variables that are categorical (likert-type)
          CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                  T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2;
   missing are all (-9);              !how the missing data are coded
   classes=class(3);                  !three class solution

ANALYSIS:
   type=mixture;
   estimator=ml;                      !Maximum likelihood estimation
   starts= 200 20;                    ! 200 starting values with 20 best optimizations

model:
%overall%

%class#1%                             !item thresholds (2) estimated for each latent class
          [T1_B01$1-T1_B15$1];
          [T1_B01$2-T1_B15$2];
%class#2%
          [T1_B01$1-T1_B15$1];
          [T1_B01$2-T1_B15$2];
%class#3%
          [T1_B01$1-T1_B15$1];
          [T1_B01$2-T1_B15$2];
PLOT:
   TYPE=PLOT1 PLOT2 PLOT3;
   SERIES IS T1_B01 - T1_B15(*);
   SAVEDATA:
 FILE IS LCA_PostProbabilities_3Classes.dat;        !File name to save posterior probabilities
   SAVE ARE CPROBABILITIES;
   OUTPUT: TECH10;
```

Graded Response Model

```
!Name of datafile
DATA:      FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;

!All the variables in the datafile
VARIABLE:   NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
          T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
          T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
          T2_B12 T2_A15 T2_B15;

!variables being used
        USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                  T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15;

!Variables that are categorical (likert-type)
          CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                  T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2;

        MISSING ARE ALL (-9);          !Indicating how missing data are coded

ANALYSIS:
       Estimator = ML;                 ! Maximum likelihood estimation
       ALGORITHM=INTEGRATION;
           LINK is LOGIT;
       STARTS = 200 20;                ! 200 starting values with 20 best optimizations

 model:
          f1 by T1_B01- T1_B15*;       ! Factor loadings are all estimated

          [T1_B01$1 - T1_B15$1*];      ! Item thresholds are all estimated
          [T1_B01$2 - T1_B15$2*];

          f1@1;                        ! Factor mean = 0 and variance = 1 for identification
          [f1@0];

Output:  STDYX;                        ! standardized solution
          Residual Tech10              ! local fit info

 PLOT:
          TYPE=PLOT1;                  ! sample descriptives
          TYPE=PLOT2;                  ! IRT-relevant curves
          TYPE=PLOT3;                  ! descriptives for theta
SAVEDATA: save = fscores;              ! save factor scores (thetas)
file is Ability_GRM.dat;               ! the file where factor scores are saved
output: stdyx; TECH1;
```

Latent Transition Analysis Model with Fixed Estimation

```
data:    FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;

VARIABLE:   NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
          T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
          T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
          T2_B12 T2_A15 T2_B15;

         USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15   T2_B01 T2_A02 T2_B04 T2_A03
                T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13 T2_B12
                T2_A15 T2_B15;

          CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
                T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11
                T2_A13 T2_B12 T2_A15 T2_B15;

   missing are all (-9);
   classes = classT1(3) classT2(3);          !there are 3 classes to be estimated at time 1 and time 2

analysis:
   type=mixture;
   estimator=ml;                             !Maximum likelihood estimation
   starts= 200 20;                           !200 starting values with 20 best optimizations

model:
%overall%
         [classT1#1];                        !intercepts of the model, one at each time point
         [classT2#1];
         classT2 on classT1;                 !regression of time 2 on time 1

Model classT1:  !assigning starting values or fixing parameters for each latent class to match
              !results from cross sectional analysis at time 1

         %classT1#1%
                  [T1_B01$1@-15 T1_A02$1@-15 T1_B04$1@-15 T1_A03$1@-15 T1_A04$1@-15
                  T1_A05$1@-15 T1_B06$1@-15 T1_B07$1@-2.911 T1_A10$1@-15
                  T1_A11$1@-1.568 T1_B10$1@-15 T1_B11$1@-15 T1_A13$1@-2.419
                  T1_B12$1@-3.512 T1_A15$1@-15 T1_B15$1@-15];

                  [T1_B01$2@-0.107 T1_A02$2@-1.971 T1_B04$2@-2.390 T1_A03$2@-2.031
                  T1_A04$2@-1.088 T1_A05$2@-2.895 T1_B06$2@-1.870 T1_B07$2@-0.011
                  T1_A10$2@-0.696 T1_A11$2@0.274 T1_B10$2@-1.717 T1_B11$2@-2.316
                  T1_A13$2@0.073 T1_B12$2@-0.851 T1_A15$2@-1.709 T1_B15$2@-14.9];

         %classT1#2%
                  [T1_B01$1@-1.366 T1_A02$1@-3.344 T1_B04$1@-15 T1_A03$1@-3.698
                  T1_A04$1@-2.37 T1_A05$1@-15 T1_B06$1@-3.425 T1_B07$1@-1.078
                  T1_A10$1@-1.144 T1_A11$1@-0.744 T1_B10$1@-2.542 T1_B11$1@-2.921
                  T1_A13$1@-2.582 T1_B12$1@-2.143 T1_A15$1@-2.075 T1_B15$1@-4.534];
```

```
        [T1_B01$2@2.688 T1_A02$2@1.241 T1_B04$2@-0.031 T1_A03$2@0.325
        T1_A04$2@0.368 T1_A05$2@0.132 T1_B06$2@0.926 T1_B07$2@2.050
        T1_A10$2@2.023 T1_A11$2@1.253 T1_B10$2@0.352 T1_B11$2@0.917
        T1_A13$2@0.779 T1_B12$2@0.939 T1_A15$2@1.040 T1_B15$2@0.084];

    %classT1#3%
        [T1_B01$1@1.636 T1_A02$1@0.018 T1_B04$1@-1.027 T1_A03$1@-0.794
        T1_A04$1@0.519 T1_A05$1@-0.560 T1_B06$1@-0.583 T1_B07$1@1.286
        T1_A10$1@0.355 T1_A11$1@0.610 T1_B10$1@-0.617 T1_B11$1@0.131
        T1_A13$1@0.070 T1_B12$1@-0.366 T1_A15$1@-0.524 T1_B15$1@-0.292];

        [T1_B01$2@4.231 T1_A02$2@4.199 T1_B04$2@2.786 T1_A03$2@3.198
        T1_A04$2@2.534 T1_A05$2@2.538 T1_B06$2@2.827 T1_B07$2@4.231
        T1_A10$2@3.785 T1_A11$2@4.217 T1_B10$2@2.439 T1_B11$2@2.769
        T1_A13$2@2.443 T1_B12$2@2.694 T1_A15$2@3.508 T1_B15$2@15];

Model classT2:          !setting item parameters in each class the same as in time 1
    %classT2#1%
        [T2_B01$1@-15 T2_A02$1@-15 T2_B04$1@-15 T2_A03$1@-15 T2_A04$1@-15
        T2_A05$1@-15 T2_B06$1@-15 T2_B07$1@-2.911 T2_A10$1@-15
        T2_A11$1@-1.568 T2_B10$1@-15 T2_B11$1@-15 T2_A13$1@-2.419
        T2_B12$1@-3.512 T2_A15$1@-15 T2_B15$1@-15];

        [T2_B01$2@-0.107 T2_A02$2@-1.971 T2_B04$2@-2.390 T2_A03$2@-2.031
        T2_A04$2@-1.088 T2_A05$2@-2.895 T2_B06$2@-1.870 T2_B07$2@-0.011
        T2_A10$2@-0.696 T2_A11$2@0.274 T2_B10$2@-1.717 T2_B11$2@-2.316
        T2_A13$2@0.073 T2_B12$2@-0.851 T2_A15$2@-1.709 T2_B15$2@-14.9];

    %classT2#2%
        [T2_B01$1@-1.366 T2_A02$1@-3.344 T2_B04$1@-15 T2_A03$1@-3.698
        T2_A04$1@-2.37 T2_A05$1@-15 T2_B06$1@-3.425 T2_B07$1@-1.078
        T2_A10$1@-1.144 T2_A11$1@-0.744 T2_B10$1@-2.542 T2_B11$1@-2.921
        T2_A13$1@-2.582 T2_B12$1@-2.143 T2_A15$1@-2.075 T2_B15$1@-4.534];

        [T2_B01$2@2.688 T2_A02$2@1.241 T2_B04$2@-0.031 T2_A03$2@0.325
         T2_A04$2@0.368 T2_A05$2@0.132 T2_B06$2@0.926 T2_B07$2@2.050
        T2_A10$2@2.023 T2_A11$2@1.253 T2_B10$2@0.352 T2_B11$2@0.917
        T2_A13$2@0.779 T2_B12$2@0.939 T2_A15$2@1.040 T2_B15$2@0.084];

    %classT2#3%
        [T2_B01$1@1.636 T2_A02$1@0.018 T2_B04$1@-1.027 T2_A03$1@-0.794
         T2_A04$1@0.519 T2_A05$1@-0.560 T2_B06$1@-0.583 T2_B07$1@1.286
        T2_A10$1@0.355 T2_A11$1@0.610 T2_B10$1@-0.617 T2_B11$1@0.131
        T2_A13$1@0.070 T2_B12$1@-0.366 T2_A15$1@-0.524 T2_B15$1@-0.292];

        [T2_B01$2@4.231 T2_A02$2@4.199 T2_B04$2@2.786 T2_A03$2@3.198
        T2_A04$2@2.534 T2_A05$2@2.538 T2_B06$2@2.827 T2_B07$2@4.231
        T2_A10$2@3.785 T2_A11$2@4.217 T2_B10$2@2.439 T2_B11$2@2.769
        T2_A13$2@2.443 T2_B12$2@2.694 T2_A15$2@3.508 T2_B15$2@15];
PLOT:
   TYPE=PLOT1 PLOT2 PLOT3;          !requests plots and post probabilities for class assignment
   SERIES IS T1_B01 - T1_B15(*);
   SAVEDATA:
  FILE IS LTA_postprob_FIXED.dat;
   SAVE ARE CPROBABILITIES;
   OUTPUT: TECH10;
```

```
data:   FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;

VARIABLE:   NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
          T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
          T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
          T2_B12 T2_A15 T2_B15;

         USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                  T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15   T2_B01 T2_A02 T2_B04 T2_A03
                  T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13 T2_B12
                  T2_A15 T2_B15;

          CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                  T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
                  T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11
                  T2_A13 T2_B12 T2_A15 T2_B15;

   missing are all (-9);

   classes = classT1(3) classT2(3);          !there are 3 classes to be estimated at time 1 and time 2

analysis:
   type=mixture;
   estimator=ml;
   starts= 200 20;

model:
         %overall%
                  [classT1#1];               !intercepts of the model, one at each time point
                  [classT2#1];
         classT2 on classT1;                 !regression of time 2 on time 1

Model classT1:    !assigning starting values for item parameters in each class that are from cross sectional
analysis results at time 1, constraining equality across time

%classT1#1%
         [T1_B01$1*-15 T1_A02$1*-15 T1_B04$1*-15 T1_A03$1*-15 T1_A04$1*-15] (1 - 5);
         [T1_A05$1*-15 T1_B06$1*-15 T1_B07$1*-2.911 T1_A10$1*-15]           (6 - 9);
         [T1_A11$1*-1.568 T1_B10$1*-15 T1_B11$1*-15 T1_A13$1*-2.419 ]       (10 - 13);
         [T1_B12$1*-3.512 T1_A15$1*-15 T1_B15$1*-15]                        (14 - 16);

         [T1_B01$2*-0.107 T1_A02$2*-1.971 T1_B04$2*-2.390 T1_A03$2*-2.031]  (17 - 20);
         [T1_A04$2*-1.088 T1_A05$2*-2.895 T1_B06$2*-1.870 T1_B07$2*-0.011]  (21 - 24);
         [T1_A10$2*-0.696 T1_A11$2*0.274 T1_B10$2*-1.717 T1_B11$2*-2.316]   (25 - 28);
         [ T1_A13$2*0.073 T1_B12$2*-0.851 T1_A15$2*-1.709 T1_B15$2*-15]     (29 - 32);

%classT1#2%
         [T1_B01$1 - T1_B15$1]                                              (33 - 48);
         [T1_B01$2 - T1_B15$2]                                              (49 - 64);

%classT1#3%
```

```
        [T1_B01$1 - T1_B15$1]                                                   (65 - 80);
        [T1_B01$2 -T1_B15$2]                                                    (81 - 96);
Model classT2:          !starting values and equality constraints to time 1

%classT2#1%
        [T2_B01$1*-15 T2_A02$1*-15 T2_B04$1*-15 T2_A03$1*-15 T2_A04$1*-15] (1 - 5);
        [T2_A05$1*-15 T2_B06$1*-15 T2_B07$1*-2.911 T2_A10$1*-15]            (6 - 9);
        [T2_A11$1*-1.568 T2_B10$1*-15 T2_B11$1*-15 T2_A13$1*-2.419]         (10 - 13);
        [T2_B12$1*-3.512 T2_A15$1*-15 T2_B15$1*-15]                         (14 - 16);

        [T2_B01$2*-0.107 T2_A02$2*-1.971 T2_B04$2*-2.390 T2_A03$2*-2.031]   (17 - 20);
        [T2_A04$2*-1.088 T2_A05$2*-2.895 T2_B06$2*-1.870 T2_B07$2*-0.011]   (21 - 24);
        [T2_A10$2*-0.696 T2_A11$2*0.274 T2_B10$2*-1.717 T2_B11$2*-2.316]    (25 - 28);
        [ T2_A13$2*0.073 T2_B12$2*-0.851 T2_A15$2*-1.709 T2_B15$2*-15]      (29 - 32);

%classT2#2%
        [T2_B01$1 - T2_B15$1]                                               (33 - 48);
        [T2_B01$2 - T2_B15$2]                                               (49 - 64);

%classT2#3%
        [T2_B01$1 - T2_B15$1]                                               (65 - 80);
        [T2_B01$2 -T2_B15$2]                                                (81 - 96);

PLOT:
        TYPE=PLOT1 PLOT2 PLOT3;
        SERIES IS T1_B01 - T1_B15(*);
         SAVEDATA:
 FILE IS LTA_postprob_JOINT.dat;
 SAVE ARE CPROBABILITIES;
 OUTPUT: TECH10;
```

Longitudinal Graded Response Model with Fixed Estimation Method

```
DATA:      FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;

VARIABLE:  NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
           T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
           T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
           T2_B12 T2_A15 T2_B15;

          USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15  T2_B01 T2_A02 T2_B04 T2_A03
                 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13 T2_B12
                 T2_A15 T2_B15;

           CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
                 T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11
                 T2_A13 T2_B12 T2_A15 T2_B15;

          MISSING ARE ALL (-9);

ANALYSIS:
         Estimator = ML;
         ALGORITHM=INTEGRATION;
             LINK is LOGIT;
         STARTS = 200 20;

 model:

! Factor loadings are all fixed based on GRM results at time 1, constrained equal across time

       f1 by T1_B01@2.28 T1_A02@2.60 T1_B04@2.40 T1_A03@2.22        (1 - 4);
       f1 by T1_A04@2.01 T1_A05@2.82 T1_B06@2.14 T1_B07@1.77        (5 - 8);
       f1 by T1_A10@1.58 T1_A11@1.08 T1_B10@1.76 T1_B11@2.50        (9 - 12);
       f1 by T1_A13@1.30 T1_B12@1.42 T1_A15@1.86 T1_B15@3.74        (13 - 16);

       f2 by T2_B01@2.28 T2_A02@2.60 T2_B04@2.40 T2_A03@2.22        (1 - 4);
       f2 by T2_A04@2.01 T2_A05@2.82 T2_B06@2.14 T2_B07@1.77        (5 - 8);
       f2 by T2_A10@1.58 T2_A11@1.08 T2_B10@1.76 T2_B11@2.50        (9 - 12);
       f2 by T2_A13@1.30 T2_B12@1.42 T2_A15@1.86 T2_B15@3.74        (13 - 16);

            f2 with f1;          !Factors covary

! Item thresholds are fixed based on GRM results at time 1, constrained equal across time

            [T1_B01$1@-0.80 T1_A02$1@-2.71 T1_B04$1@-3.9]        (17 - 19);
            [T1_A03$1@-3.33 T1_A04$1@-1.68 T1_A05$1@-3.74]       (20 - 22);
            [T1_B06$1@-2.97 T1_B07$1@-0.61 T1_A10$1@-1.10]       (23 - 25);
            [T1_A11$1@-0.50 T1_B10$1@-2.48 T1_B11$1@-2.4]        (26 - 28);
            [T1_A13$1@-1.57 T1_B12$1@-1.88 T1_A15$1@-2.31]       (29 - 31);
            [T1_B15$1@-4.16]                                     (32);

            [T1_B01$2@3.22 T1_A02$2@1.91 T1_B04$2@0.55]          (33 - 35);
```

```
        [T1_A03$2@0.90 T1_A04$2@0.99 T1_A05$2@0.70]      (36 - 38);
         [T1_B06$2@1.34 T1_B07$2@2.54 T1_A10$2@2.09]     (39 - 41);
        [T1_A11$2@1.73 T1_B10$2@0.75 T1_B11$2@1.42]      (42 - 44);
        [T1_A13$2@1.36 T1_B12$2@1.26 T1_A15$2@1.39]      (45 - 47);
        [T1_B15$2@0.94]                                  (48);

        [T2_B01$1@-0.80 T2_A02$1@-2.71 T2_B04$1@-3.9]    (17 - 19);
        [T2_A03$1@-3.33 T2_A04$1@-1.68 T2_A05$1@-3.74]   (20 - 22);
        [T2_B06$1@-2.97 T2_B07$1@-0.61 T2_A10$1@-1.10]   (23 - 25);
        [T2_A11$1@-0.50 T2_B10$1@-2.48 T2_B11$1@-2.4]    (26 - 28);
        [T2_A13$1@-1.57 T2_B12$1@-1.88 T2_A15$1@-2.31]   (29 - 31);
         [T2_B15$1@-4.16]                                (32);

        [T2_B01$2@3.22 T2_A02$2@1.91 T2_B04$2@0.55]      (33 - 35);
        [T2_A03$2@0.90 T2_A04$2@0.99 T2_A05$2@0.70]      (36 - 38);
        [T2_B06$2@1.34 T2_B07$2@2.54 T2_A10$2@2.09]      (39 - 41);
        [T2_A11$2@1.73 T2_B10$2@0.75 T2_B11$2@1.42]      (42 - 44);
        [T2_A13$2@1.36 T2_B12$2@1.26 T2_A15$2@1.39]      (45 - 47);
        [T2_B15$2@0.94]                                  (48);

        f1@1;                       ! Factor mean = 0 and variance = 1 for identification
        [f1@0];
        f2@1;
        [f2@0];

Output: STDYX;                      ! standardized solution
        Residual Tech10             ! local fit info

PLOT:
        TYPE=PLOT1;                 ! sample descriptives
        TYPE=PLOT2;                 ! IRT-relevant curves
        TYPE=PLOT3;                 ! descriptives for theta

  SAVEDATA: save = fscores;         ! save factor scores (thetas)
  file is Ability_LONG_FIXED.dat;   ! the file where factor scores are saved

output: stdyx; TECH1;
```

```
DATA:      FILE IS FINAL_Longitudinal_AnalysisData_MPLUS.dat;
VARIABLE:  NAMES ARE ClientID T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06
           T1_B07 T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
           T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13
           T2_B12 T2_A15 T2_B15;

         USEV = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07 T1_A10 T1_A11
                T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15   T2_B01 T2_A02 T2_B04 T2_A03
                T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11 T2_A13 T2_B12
                T2_A15 T2_B15;

          CATEGORICAL = T1_B01 T1_A02 T1_B04 T1_A03 T1_A04 T1_A05 T1_B06 T1_B07
                T1_A10 T1_A11 T1_B10 T1_B11 T1_A13 T1_B12 T1_A15 T1_B15 T2_B01 T2_A02
                T2_B04 T2_A03 T2_A04 T2_A05 T2_B06 T2_B07 T2_A10 T2_A11 T2_B10 T2_B11
                T2_A13 T2_B12 T2_A15 T2_B15;

         MISSING ARE ALL (-9);

ANALYSIS:
        Estimator = ML;
        ALGORITHM=INTEGRATION;
            LINK is LOGIT;
        STARTS = 200 20;

 model:
          f1 by T1_B01 - T1_B15*        (1 - 16); ! Factor loadings are estimated – equal across time
          f2 by T2_B01 - T2_B15*        (1 - 16);
           f2 with f1;                          !Factors covary

          [T1_B01$1 - T1_B15$1]         (17 - 32); ! Item thresholds are estimated – equal across time
          [T1_B01$2  - T1_B15$2]        (33 - 48);
          [T2_B01$1 - T2_B15$1]         (17 - 32);
          [T2_B01$2 - T2_B15$2]         (33 - 48);

          f1@1;                                 ! Factor mean = 0 and variance = 1 for identification
          [f1@0];
          f2@1;
          [f2@0];

Output:   STDYX;                                ! standardized solution
          Residual Tech10                       ! local fit info
 PLOT:
          TYPE=PLOT1;                           ! sample descriptives
          TYPE=PLOT2;                           ! IRT-relevant curves
          TYPE=PLOT3;                           ! descriptives for theta
   SAVEDATA: save = fscores;                    ! save factor scores (thetas)
  file is Ability_LONG_JOINT.dat;               ! the file where factor scores are saved
output: stdyx; TECH1;
```