EXPLORING THE CONSTRUCT VALIDITY OF TWO IRT-DERIVED SCALES OF THE PAI-A

A Thesis

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Master of Arts

By

Amelia D. Coffman

May, 2013

EXPLORING THE CONSTRUCT VALIDITY OF TWO

IRT-DERIVED SCALES OF THE PAI-A

_____

Amelia D. Coffman

APPROVED:

_____

Lynne Steinberg, Ph.D.
Committee Chair

_____

Carla Sharp, Ph.D.

_____

Leslie C. Morey, Ph.D.
Department of Psychology
Texas A&M University

_____

John W. Roberts, Ph.D.
Dean, College of Liberal Arts and Social Sciences
Department of English

EXPLORING THE CONSTRUCT VALIDITY OF TWO IRT-DERIVED SCALES OF THE PAI-A

An Abstract of a Thesis

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Master of Arts

By

Amelia D. Coffman

May, 2013

ABSTRACT

The Personality Assessment Inventory – Adolescent (PAI-A) is a 264-item self-report instrument designed to assess various facets of psychopathology and of personality in adolescents. Three studies were used to examine the performance of the Anxiety (ANX) and Depression (DEP) scales of the PAI-A, the first two studies using item response theory (IRT) methods to revise the ANX and DEP scales of the PAI-A using the graded response model (Samejima, 1996), and the third study investigating the construct validity of these IRT-derived scales. In Study 1, item performance was examined separately for each scale in a community sample (N = 707) of adolescents; Study 2 examined the PAI-A ANX and DEP scales in a clinical sample of adolescents (N = 1160). On the basis of these IRT analyses, some items were eliminated from the ANX and DEP scales, resulting in 13-item sets for both scales. In Study 3, the construct validity of the IRT-derived item sets as well as the original full-length ANX and DEP scale scores was investigated using a sample of adolescents admitted to the adolescent program of a private tertiary care inpatient treatment facility (N = 169). Correlations were computed between the PAI-A full-length and the IRT-derived ANX and DEP scales and each continuously scored validating instrument. These instruments included the anxiety and depression scale scores of the Youth Self-Report (YSR), Child Behavior Checklist (CBCL/6-18), Multidimensional Anxiety Scale for Children (MASC), and the Beck Depression Inventory-II (BDI-II). Correlations were strongest for the two most narrowly focused validating measures (i.e., MASC and BDI-II), and were notably smaller for parent-report measures (i.e., P-DISC and CBCL/6-18). For the categorically scored Youth and Parent Diagnostic Interview Schedule for Children (Y-DISC and P-DISC), *t*-tests of PAI-A

scores between DISC diagnostic groups were conducted. For the full-length PAI-A scale scores, those with a positive diagnosis on the youth and parent DISC had higher scores compared to those with no/intermediate diagnosis. For the IRT-revised scores, mean differences were found between the youth DISC categories for ANX and DEP; however, differences between the parent DISC categories were found only for DEP. Those with a diagnosis had higher scores compared to those with no/intermediate diagnosis.
Keywords: anxiety, depression, PAI-A, item response theory, graded response model

ACKNOWLEDGEMENTS

I want to extend lasting thanks to my committee, Drs. Lynne Steinberg, Carla Sharp, and Les Morey. Thank you all for your guidance and patience throughout this project. I also thank Dr. Carla Sharp and Dr. Les Morey for their individual contributions of data that made this research possible.

My deep appreciation goes to both Dr. Les Morey and Dr. Terence Hoagwood for their direction as I first considered beginning my journey into graduate school. I am also particularly thankful to Dr. Hye Jeong Kim for her support throughout much of my higher education.

I would like to thank my parents and family for supporting me without question. You have all tolerated my thinking and writing and have given me many opportunities to work. Mom and Dad, thank you for never tiring of my endless thesis chatter, and for your continual faith in me.

I appreciate all of you immensely.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

I. TABLES

I dedicate these pages with love to the memory of

L.L. and Audrie Lee Coffman

and

Kermitt H. Marshall

Exploring the Construct Validity of Two IRT-Derived Scales of the PAI-A

The accurate assessment of adolescent depression and anxiety symptomology is both important and challenging. Due to the frequency of occurrence for both disorders across populations (Abela & Hankin, 2008; Dobson & Cheung, 1990; Esbjørn, Hoeyer, Dyrborg, Leth, & Kendall, 2010), the rapid increase of depression and anxiety in adolescence (Dierker et al., 2001; Fichter, Kohlboeck, Quadflieg, Wyschkon, & Esser, 2009; Pine, Cohen, Gurley, Brook, & Ma, 1998), and the ability of adolescent anxious and depressive symptoms to predict a large variety of adulthood psychopathology (Copeland, Shanahan, Costello, & Angold, 2009; Costello, Copeland, & Angold, 2011; Devine, Kempton, & Forehand, 1994; Keenan, Feng, Hipwell, & Klostermann, 2009), the adolescent assessment of these disorders has the potential to be helpful to individuals across stages of life. This study assesses the Anxiety and Depression subscales of the Personality Assessment Inventory – Adolescent (PAI-A; Morey, 2007) using item response theory (IRT) methods. IRT will be used to evaluate the measurement models and item performance, and can be expected to result in a reduced number of items in these two PAI-A scales. To culminate this study, construct validity of the two IRT-derived scales will be assessed using data collected from multiple instruments as part of an adolescent inpatient treatment program.

The adult Personality Assessment Inventory (PAI; Morey, 1991) is a 344-item self-report instrument designed to assess various facets of psychopathology and of personality. Its adolescent counterpart, the Personality Assessment Inventory – Adolescent (Morey, 2007), is a 264-item form of the original instrument. Both instruments share the same scale structure of 22 nonoverlapping scales: 4 validity scales

that indicate random or negatively/positively skewed responding, 11 clinical scales that each address an area of psychopathology, 5 treatment consideration scales (e.g, Aggression, Stress), and 2 interpersonal scales that correspond to the axes of the interpersonal circumplex (Dominance and Warmth; Leary, 1957). This study focuses exclusively on two of the PAI-A's clinical scales, Depression (DEP) and Anxiety (ANX).

Development of the PAI & PAI-A

The initial development of the PAI came at a crucial time in the assessment of psychopathology. Together with the PAI, the revised Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), and the revised Millon Clinical Multiaxial Inventory (MCMI-III; Millon, 1994), make up the three multiscale measures of personality and psychopathology most commonly in use today (Davis & Archer, 2010; Wise, Streiner, & Walfish, 2010). There are multiple theoretical and psychometric limitations evident in the widely popular MMPI-2 (cf. Helmes & Reddon, 1993; Horvath, 1992; Wetzler & Marlowe, 1992). In spite of the numerous empirical strengths of the MMPI-2, concerns such as the applicability of norming groups and the inclusion of items that may no longer be relevant, in addition to a lack of any explicit theoretical approach have caused many to question the current utility of the MMPI-2 (Helmes & Reddon, 1993, give a complete critical review). In contrast, the MCMI-III has the benefit of being more strongly theory driven, but in its earlier revisions it lacked the extensive validity research which supports the MMPI-2 (Dana & Cantrell, 1988). For both the MMPI-2 and MCMI-III, item overlap between scales may create a major problem (Anderson & Bashaw, 1966; Helmes & Reddon, 1993; Hsu, 1994, 2005).

One of the motivations behind the development of the PAI seems to have been the opportunity to address some of these areas of concern. The PAI has received praise for its development using both theoretical and empirical methods (Hopwood, Blais, & Baity, 2010), for its resolution of a number of psychometric shortcomings of previous instruments (Helmes, 1993), and for using clearer and more clinically modern concepts (Schlosser, 1992), among other detailed improvements mentioned by these authors. Adolescent versions of each of the above-mentioned instruments have also been released, and many of the concerns regarding the adult instruments are still evident in these adolescent measures (Merydith & Phelps, 2009), making both the PAI and PAI-A valuable and needed contributions to the field of multi-scale psychopathology instruments.

The development of the clinical scales of the adult PAI followed a construct validation approach consisting of theory formulation, internal validation, and external validation (Morey, 1991, 2000; Morey & Hopwood, 2006, 2008; Morey, Warner, & Hopwood, 2007). The DSM-III (American Psychiatric Association, 1987), in use during the creation of the PAI, documents numerous disorders that were examined for possible inclusion in the clinical scales of the PAI. Morey (1991) describes two specific selection criteria for the PAI's clinical scales: (1) a disorder's stability over time, and (2) a disorder's significance in clinical practice. The consideration of potential disorders for the PAI resulted in the eventual inclusion of eleven specific clinical scales: Somatic Complaints, Anxiety, Anxiety-Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, and Drug Problems. The two scales used in this study, Depression and Anxiety, were selected for

this research for several reasons. Depression and anxiety are arguably the most well researched, both separately and jointly, and the most correlated of any two scales (Ingram & Siegle, 2009; Moran & Lambert, 1983). These two clinical scales share the same subscale structure in both the PAI and PAI-A (cognitive, affective, and physiological subscales), which makes the pair well-suited for this research. Moreover, there are additional accepted and well-researched measures for each of these concepts (Nezu, Nezu, Friedman, & Lee, 2009), allowing for a thorough exploration of construct validity. Additionally, these two constructs represent two of the more widely experienced clinical conditions in adolescents (Abela & Hankin, 2008), allowing for greater potential application of the conclusions of this research.

Two shortened versions of the PAI have been released: the PAI-Short Form (PAI-SF) and the Personality Assessment Screener (PAS). The PAI-SF is comprised of the first 160 items of the PAI, where these first items were chosen based on their highest item-scale correlations. The PAI-SF ANX and DEP scales consist of 12 items each (Morey, 1991). The PAS is a 22-item instrument designed to screen respondents for the need for full evaluation with the PAI. The PAS has ten elements, but has no separate, specific scales for anxiety or depression (Creech, Evardone, Braswell, & Hopwood, 2010). The reduction in items for the PAI-SF and the PAS were not created using item response theory methods.

Anxiety. Anxiety disorders in adolescence are consistently reported as being quite pervasive, though the exact reported lifetime prevalence for adolescents varies from approximately 10% – 20% (Weems & Silverman, 2008). In addition to these high prevalence rates, there is wide agreement that child and adolescent anxiety disorders are

highly predictive of adult psychopathology (Abela & Hankin, 2008; Pine, et al., 1998), of both homotypic (other anxiety-spectrum disorders) and heterotypic (non-anxiety symptomology) continuity (Esbjørn, et al., 2010).

Anxiety is often defined either in terms of different anxiety disorders taken separately (Social Anxiety Disorder, Panic Disorder, etc.) or in terms of generalized anxiety, a broader conceptualization of anxiety that is not focused on a specific object or situation. The ANX scale of the PAI-A focuses on this generalized anxiety. Several potential elements of anxiety are proposed, with possible components generally including cognitive, behavioral, somatic/physiological, and affective concerns (Koksal & Power, 1990; Weems & Silverman, 2008). Due to the difficulty in measuring the behavioral component of anxiety uniformly across various specific anxiety behaviors (Kearney & Bensaheb, 2007; Morey, 1991, 2007), this component was omitted from the ANX subscale of the PAI, and the behavioral component of anxiety was included in the PAI and PAI-A as a separate scale, Anxiety-Related Disorders. This subscale will not be included in the present study.

Morey (1991) describes the content covered by each of the three Anxiety subscales as: (1) Cognitive Anxiety (ANX-C), composed of "cognitive beliefs, expectation of harm, ruminative worry," (2) Affective Anxiety (ANX-A), made up of "feelings of tension, panic, nervousness, " and (3) Physiological Anxiety (ANX-P), which includes somatic complaints such as "racing heart, sweaty palms, rapid breathing, and dizziness."

Depression. Depression in adolescents is consistently reported as a predictor of future psychopathology (Klein, Torpey, & Bufferd, 2008) and of impairment across

domains of functioning (Abela & Hankin, 2008; Rudolph, 2009). Along with anxiety disorders, rates of depressive disorders begin to increase rapidly during adolescence, with depressive symptoms becoming particularly apparent around ages 13 – 15 (Rudolph, 2009), making the problems presented by depression especially salient for adolescents (Klein, et al., 2008).

The PAI-A DEP scale was created with the scope of existing depression scales in mind, and with the observation that many depression scales have a somewhat focused concentration (e.g., the Beck Depression Inventory has a largely cognitive focus; Morey, 1996; 2000). Depression is often conceptualized using the same structure as anxiety (Moran & Lambert, 1983), with the three PAI-A DEP subscales of *cognitive*, *affective*, and *physiological*.

Morey (1991) describes the content covered by each of the three Depression subscales as: (1) Depression-Cognitive (DEP-C), focusing mainly on "thoughts of worthlessness, hopelessness, and personal failure," (2) Depression-Affective (DEP-A), made up of "feelings of sadness, loss of interest in activities, and anhedonia," and (3) Depression-Physiological (DEP-P), which includes somatic disturbances such as "disturbance(s) in sleep pattern and changes in appetite."

Transition to the PAI-A

While the PAI has occasionally been used in adolescent populations, the need for an adolescent version of the instrument led to the development of the PAI-A, designed for use with adolescents ages 12-18 (Morey, 2007). While many current adolescent measures of psychopathology are developed as "downward extensions" of adult measures (Kearney & Bensaheb, 2007), the PAI-A is unique in that it not only retains the same

item and scale structure as the adult PAI, but also that it preserves the meaning of items from the perspective of the adolescent test-taker. Morey (2007) examined PAI items and worked to confirm that the items selected for the PAI-A retained their original (PAI) meaning when read by an adolescent, for example, an item's only difference between the adult and adolescent instrument might be the changing of "work" to "school." The PAI-A beta version (all 344 PAI items) was administered to a group of 275 adolescents and comparisons with the adult PAI were examined. Items that appeared to perform differently on the PAI and the PAI-A beta were eliminated in most cases (Morey, 2007). This allowed for a shorter adolescent version of the test and also attempted to prevent any unforeseeable effects caused by the rewording of test items. While the PAI contains more items in all of its scales, the PAI-A does retain the scale structure and response options of the adult instrument. For example, the PAI ANX scale has 24 items and the PAI-A ANX scale has only 18 items, but all of the adolescent items are equivalent to adult items. In fact, for the DEP and ANX scales, the PAI-A shares exact item wording with PAI items (though the adult measure is longer and thus contains additional items not represented in the PAI-A). Along with PAI-A research, studies focusing exclusively on the use of the PAI with adult participants will be considered as part of this review, and based on the similarity of scales and items between the adult and the adolescent instrument, will help inform knowledge of the adolescent PAI-A.

Psychometric Properties of the PAI and PAI-A: Review of Research

The PAI and PAI-A have been used with many samples in multiple settings, including with chronic pain patients (Hopwood, Creech, Clark, Meagher, & Morey, 2007), athletes (Storch, Storch, Killiany, & Roberti, 2005), patients admitted due to

substance abuse (Hopwood, Baker, & Morey, 2008), forensic samples (Morey, et al., 2007), predictions of client-initiated therapy termination (Hopwood, Ambwani, & Morey, 2007), samples with psychopathy (Blonigen et al., 2010), inpatient samples (Siefert, Sinclair, Kehl-Fie, & Blais, 2009), and samples with eating disorders (Tasca, Wood, Demidenko, & Bissada, 2002).

Several research studies, discussed in the following paragraphs, have investigated the psychometric properties of the PAI, and, to a lesser extent, the PAI-A. When exploring available literature on the PAI and PAI-A, it is evident that most research studies have focused either on special populations (e.g., incarcerated adolescents), or on specific scales (e.g., only the Antisocial Features scale). Additionally, some studies (before the 2007 release of the PAI-A) use the adult PAI in adolescent populations (e.g., Hoekstra, 2000). This review will first address scale-level factor analytic studies, followed by item-level analyses, finally concluding with existing IRT research.

Some similarities lie specifically within the DEP and ANX scales used in this research. Depression and anxiety are frequently seen in a similar light, whether by their conceptual similarity (Dobson & Cheung, 1990), their same internalizing characteristics (Ruiz & Edens, 2008), or the frequent comorbidity between the two (Clark & Watson, 1991; Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Rudolph, 2009; Semrud-Clikeman, Fine, & Butcher, 2007). Factor analyses of the PAI subscale scores have invariably placed DEP and ANX together on a single factor. This was true whether the analyses used all 22 PAI scales or only the 11 clinical scales, and across factor analytic methods. For example, in a factor analysis of all 22 scales of the PAI-A , Morey (2007) found that one factor (factor one in the professional manual) seemed to represent general

psychological distress, and included both ANX and DEP (along with other scales). This factor in particular has been well replicated for the PAI: other factor analyses, discussed below, have duplicated this pairing of Depression and Anxiety on the same factor, even when the remaining factor structure differed slightly.

While the current research focuses on the item-level psychometric properties of the DEP and ANX scales, most PAI and PAI-A studies have focused on very specific populations of clinical interest. Of the more psychometrically-oriented studies, the bulk of research focuses on scale-level rather than item-level statistics. In scale-level research, the most notable studies are, in fact, those that explore the factor structure of the PAI, such as studies by Boyle and colleagues (Boyle & Lennon, 1994; Boyle, Ward, & Lennon, 1994), Deisinger (1995), and Hoelzle and Meyer (2009).

Boyle and Lennon's (1994) examination of the PAI aimed to replicate Morey's (1991) reported four factors (most notable here is Morey's first factor, which includes ANX and DEP). Two factor analyses were conducted using all 22 PAI scales: the initial analysis was conducted on an Australian sample of nonclinical participants (n = 151), participants admitted to an inpatient unit for alcoholism (n = 30), and participants admitted for schizophrenia (n = 30), and a second analysis used Morey's (1991) clinical standardization data (N = 1246). While Boyle and Lennon criticize the factor analytic approach reported in the PAI professional manual (notably the orthogonal varimax rotation), both analyses do produce a factor much like the first factor reported in the PAI's professional manual (though there are some differences in the remaining factor structure).

Deisinger (1995) also used factor analysis to explore the structure of the PAI, with

a set of factor analyses conducted on the PAI's full 22 scales as well as on its 11 clinical scales. The sample (N = 168) was nonclinical and completed the PAI as a part of a related study. Deisinger used an oblique rotation due to the probable correlation between many of the clinical scales (and comorbidity between many of the core constructs), though the factor solutions obtained were very similar to those reported by Morey (1991). In the 22-scale analysis, a four-factor solution closely matched the four factors Morey reported. The 11 clinical scales fit into a four-factor model, but were best explained by a three-factor solution. Still, this three-factor model includes a factor very similar to Morey's first factor (with strong loadings for DEP and ANX; eigenvalues both > .90). Thus, while the remaining factor structure differed to some degree, especially in the 11 clinical scales, the factor representing "general distress" appears to be relatively consistent across samples and factor analysis strategies.

Hoelzle and Meyer (2009) reviewed 11 articles containing 21 separate factor analytic research studies (including those discussed above and the factor analyses reported in the PAI professional manual), again uniformly finding DEP and ANX on one factor. Of these 21 factor analyses, 11 use specialized samples unrelated to the present research. The remaining 10 studies include those already discussed: those in the PAI professional manual (Morey, 1991), those by Boyle and Lennon (1994), and those by Deisinger (1995). This detailed comparison of individual methods (11 or 22 scales analyzed, inclusion/exclusion criteria, type of factor analysis performed, type of rotation, and criteria for extraction) is at the scale (not item) level, making these results interesting, but not directly applicable to this more tightly focused, item-level investigation of ANX and DEP.

While factor analytic studies provide some information about the global structure of the PAI, item-level analyses of the PAI or PAI-A are needed to help inform the functioning of the PAI-A at the level of the individual items. These item-level analyses are sparse, however two adult PAI studies are available. Siefert and colleagues (2009) conducted an item-level examination of the PAI in an inpatient sample (N = 646) and found support for the specific item-scale pairings used in the PAI. All PAI items were examined for cross-correlations, or unintended correlations with other PAI scales. The mean item-scale correlations and interitem correlations for ANX and DEP (and their subscales) were reasonably high. To examine the "scaling success" of an item, Siefert and colleagues (2009) also tested the hypothesis that a given item would correlate better with its intended scale than any other scale; this hypothesis was tested using a *t*-test for dependent correlations with the results that, for ANX and DEP, items did correlate best with their intended scale. The cross-scale correlations, or mismatched item-scale pairs, were not significant for both scales.

Only one study was found that analyzed either the PAI or PAI-A using IRT methods. In his dissertation, Gouge (2009) examined all 22 PAI scales of the PAI at the item level using IRT (parametric and nonparametric IRT). Though the initial goal of his study is not unlike the goal of the present research, the population studied was methadone maintenance treatment patients, which may limit the generalizability of the findings. Additionally, Gouge conducted his analyses separately on each of the three subscales (i.e., cognitive, affective, and physiological components), rather than at the scale level. Because the current research will use the scale level to conduct analyses, the approach of Gouge's study is quite different.

Purpose of This Research

The purpose of this research is to examine and revise the ANX and DEP scales of the PAI-A using IRT methods and to study evidence for the construct validity of both the original and revised scales. This will be accomplished with three studies, the first two using IRT methods to revise the DEP and ANX scales of the PAI-A and the third investigating the construct validity of the original and the IRT-derived scales. While the PAI and PAI-A are recognized as improvements over previously existing instruments (Helmes, 1993; Helmes & Reddon, 1993; Hopwood, et al., 2010; Schlosser, 1992), and have been well-researched, research using IRT methods is lacking. This item-level IRT analysis of the DEP and ANX scales is expected to offer a key theoretical and practical advantage. Because IRT methods work at the item level, they are an appropriate tool for identifying items that may not be performing well.

This IRT analysis is thus expected to produce item-level information that will lead to better understanding of the scale as a whole and may yield a somewhat shortened scale that will reduce strain on both respondents and clinicians. After this initial IRT analysis, the construct validity of IRT-derived scales will be examined by investigating the relationships between the original PAI-A scales and the reduced scales in terms of correlations with the other measures of depression and anxiety. Hypotheses of the construct validation portion of this research are that measures of depression that correlate with the original PAI-A ANX or DEP scale will replicate these correlations in the IRT-derived PAI-A ANX or DEP scale, and likewise that significant mean differences between groups in categorical measures for the original PAI-A ANX or DEP scale will be replicated in the IRT-derived PAI-A ANX or DEP scale.

Measurement Models

Item Response Theory

IRT has now become widespread for many types of item analyses, and will provide the basis for evaluating the items of the DEP and ANX scales of the PAI-A. While a rich history underlies current IRT understanding and practices (cf. Bock, 1997; Thissen & Steinberg, 2009), some basic ideas are at the foundation of all IRT models and analyses. First, there is an assumption that there is some continuous unobserved "ability" or "trait." This underlying construct or latent trait is often termed *theta*, and is placed on a standard unit scale ranging from -3 to +3. Thus, for these analyses this latent trait will be either *anxiety* or *depression*. Additionally, the parameters of items place them along the same scale as this latent trait. One of the assumptions critical to unidimensional IRT is the need for all observed relationships among the item responses to be completely accounted for by the latent variable (Lazarsfeld, 1950, as cited in Thissen & Steinberg, 2009). This is often termed *local independence*. Additionally, for each scale the dimensionality of the item set will be evaluated. It is anticipated that *unidimensional* IRT will be used, that is, a single construct will account for the item covariation for the DEP and ANX scales, respectively.

IRT analysis relies on the ability of various mathematical models to understand the item response data. While numerous models exist for multiple category items (discussion of many of these models can be found in Embretson & Reise, 2000), the graded response model will be used in these analyses.

The graded response model. The graded response model or GRM (Samejima, 1969, 1996, 2010) is an extension of the two-parameter logistic model (2PL) for

polytomous data that is particularly well suited for Likert-type data. The GRM essentially breaks the polytomous data into binary pieces, and fits a 2PL model for the probability of a response in a category or higher (e.g., probability a response is in category 2 or higher). The GRM estimates two types of parameters for each item: one slope (a) and one less threshold (b_i) than the number of response categories (e.g., for 4 response categories, 3 thresholds are estimated). The slope, or discrimination, parameter is positive and indicates the strength of the relationship between the item response and the underlying construct. The threshold parameter provides information about the "difficulty" of endorsing an item by indicating the position on the latent variable to which the probability of a response in a category or higher crosses 50%. A higher threshold indicates that a respondent needs to be higher on the latent trait to endorse that category or higher. Thus, the value of the threshold parameter is the point at which the probability passes 0.5 that a response in category k or higher will be endorsed. The probability that a response is in a particular category is calculated from the probability that a given category k or higher is observed minus the probability of the next highest category or higher. In addition to item parameters, graphical representations of the probability of different response options, called trace lines, are helpful in describing the results of the item analysis. Test information curves will provide information about the overall functioning of each scale.

Information about the functioning of each item and its response alternatives can be interpreted from the values of the item parameters, the slope (a) and the three thresholds (b_i), and their corresponding trace lines. Items with questionable performance, that is, items that do not appear strongly related to the underlying construct will be

removed to create the revised forms of the ANX and DEP scales.

Measure

Personality Assessment Inventory – Adolescent. The Personality Assessment Inventory – Adolescent (Morey, 2007) is a 264-item self-report instrument assessing psychopathology and personality that contains 22 scales, 11 of them assessing individual clinical constructs. The PAI-A is written at a 4th-grade reading level, and can be completed in approximately 45 minutes (Morey, 2007). PAI-A respondents answer items on a four-point rating scale (not true, slightly true, mainly true, very true). The PAI-A is designed with mean *t*-scores of 50 (*SD* = 10). Based on the values reported in the professional manual, DEP and ANX show acceptable internal consistency (Morey, 2007). For DEP, coefficient alpha for the community sample is .86 and for the clinical sample is .88. For ANX, these values are .86 (community) and .89 (clinical).

Method

Using IRTPRO (Cai, du Toit, & Thissen, 2011a), data from the ANX and DEP scales of the PAI-A were studied using item parameters generated within the GRM. This included a separate analysis of the community sample (Study 1) and the clinical sample (Study 2) of the standardization data for the PAI-A (each conducted separately for ANX and DEP). Items were evaluated with attention to how they performed in relation to the underlying construct. These evaluations were based multiple criteria, described below.

Criteria for Item Selection

The general goal of Studies 1 and 2 is to evaluate the ANX and DEP items within an IRT framework. This includes maximizing the inclusion of items that relate most directly to the underlying construct, while potentially removing some less ideal items in

order to create final item sets that display minimal error and redundancy in item content. The explicit goal is not item reduction, though this is likely, but to maximize information gained from each final item set.

All response frequencies were adequate for IRT analysis. There were at least 9 responses in each response category for each item in the community sample (Study 1). In the clinical sample (Study 2), there were at least 70 responses in each response category. For both samples, each response was recorded as: 0 = not true, 1 = slightly true, 2 = mainly true, 3 = very true. Both scales contain 18 items comprised of *affective*, *cognitive*, and *physiological* subscales, each containing six items.

For the purpose of this study, several initial criteria for item inclusion were used. These criteria, discussed more thoroughly below, include: (1) absence of local dependence, (2) acceptable slope parameters, (3) clear and direct item content, (4) an awareness of threshold parameters, and (5) satisfactory final model fit.

As previously mentioned, it is expected that unidimensional IRT will be used for these analyses. A primary assumption in unidimensional IRT models is local independence; the inability to meet this assumption results in *local dependence* (LD). A LD index is calculated for each item pair, and the presence of excessive LD in some number of item pairs indicates that there is covariation greater than can be accounted for by the unidimensional model (anxiety or depression). LD is often caused by some degree of redundancy in item content. LD is also related to the assumption of unidimensionality: a measure with patterns of LD may indicate additional dimensions in the scale. LD may also cause the biased estimation of slope and thresholds, leading to incorrect parameter estimates (Yen, 1993).

LD has been measured in a number of ways over time (see Chen & Thissen, 1997; Yen, 1993), but IRTPRO software, used in these studies, makes use of Chen and Thissen's (1997) LD χ^2 index and its extension to multiple category items. The standardized LD χ^2 index is represented similarly to a chi-squared distribution, and is derived from comparing observed and expected frequencies of responses. Values are reported in a pair-wise matrix of values, where values greater than 10 indicate probable LD (Cai, Du Toit, & Thissen, 2011b).

LD is often remedied by removing one item in an LD pair, usually the item displaying low slope or unclear item wording. One other issue in interpreting the LD χ^2 values are "clusters" of LD values, or groups of high LD. These patterns lead to a questioning of the unidimensionality of the instrument, and are of special concern to the current analyses because of the expected subscale structure of ANX and DEP (affective, cognitive, physiological). That is, a unidimensional model could exhibit poor fit due to relatedness (evidenced as high LD) within items on the same subscale.

Each item's slope or discrimination parameter (a) increases as the item is more strongly related to the underlying construct, as defined by the items in the analysis. Low slope indicates that an item is measuring something less related to this underlying construct, and also produces threshold parameters that do not distinguish well between response categories. In general, a slope should not fall below 1.0, though the confidence interval for the slope may be useful in some decision-making in order to take a more conservative approach to item removal.

Threshold or difficulty parameters (b_i) are somewhat less critical than LD and slope, but may still affect item decisions. Because threshold parameters are responsible

for the horizontal shift in the response categories, having all threshold parameters roughly centered on the same area of the underlying construct effectively creates an instrument that only distinguishes respondents in a small range of that construct. For example, if all items with lower threshold parameters were removed (for any reason) in the revision process, the instrument would no longer be able to accurately discriminate respondents with these lower levels of the underlying construct: information would have been lost.

Finally, the item sets selected based on the analyses in Study 1 and Study 2 must match, as only one item set will be used for each scale in Study 3. In the event of a disagreement between community and clinical samples, items that perform well in the community sample will be retained. This decision was based on the fact that the clinical sample was very heterogeneous in terms of diagnosis, rather than containing only anxiety and depression diagnoses.

A proposed final item set for inclusion in Study 3 thus should be comprised of items that do not exhibit LD, that have sufficient slope parameters, and that, wherever possible, retain sufficient breadth of difficulty in the items. Finally, this proposed item set must exhibit close fit to the IRT model. The M_2 goodness-of-fit statistic (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2006) is helpful, though it is quite likely that there is a significant difference for many models since M_2 tests the *exact* fit of the model. As an addition to the use of the M_2 goodness-of-fit statistic, a fit statistic such as the root mean square error of approximation (RMSEA) will be used (Steiger & Lind, 1980, as cited in Browne & Cudeck, 1993). Fit indices measure the lack of fit of a model, and as such, the RMSEA can be a useful tool. The RMSEA guidelines suggested by Browne and Cudeck (1993) will guide this research. They recommend that

values of less than 0.05 represent close fit, values less than 0.08 represent a reasonable fit, and values greater than 0.1 represent poor fit. The use of the RMSEA will thus provide three fit indexes: LD, M_2, and RMSEA. Together these indices will help inform the adequacy of the model.

Study 1: Community Sample

PAI-A Community Standardization Sample

The PAI-A community sample of adolescents is a subset of the data collected for the standardization of the PAI-A, and was used with permission. These data were obtained using a stratified sampling approach based on the 2003 U.S. Census (Morey, 2007). The PAI-A was completed by 1,032 adolescents, and profiles were selected for inclusion in the final dataset based on census-data fit for age, sex, and ethnicity. The final community standardization sample (N = 707, 51.1% male) represents age and sex equally, and represents ethnicity at a census-matched frequency (see Table 1).

Table 1

Demographics of the PAI-A Community Sample of Adolescents

Age			Ethnicity		
Age	*n*	Percent	Ethnicity	*n*	Percent
12	102	14.4	White	435	61.5
13	100	14.1	African-American	109	15.4
14	101	14.3	Hispanic	115	16.3
15	101	14.3	Other	48	6.8
16	101	14.3			
17	101	14.3			
18	101	14.3			

Results

Anxiety

An item analysis using the GRM was performed on all 18 ANX items using IRTPRO; item parameters for this item set are presented in Table 2. Several instances of LD needed to be addressed, though the LD did not seem to follow a pattern associated with the items on the three subscales. The items were therefore analyzed together at the scale level rather than at the subscale level. Items 83, 110, and 123[1] all exhibited LD pairs with one another, necessitating the removal of two of these items (83 and 110 LD χ^2 = 10.1; 83 and 123 LD χ^2 = 11.6; 110 and 123 LD χ^2 = 23.7). These items all seem to relate to content concerning being relaxed. Item 110 had the largest slope of these three items, and seemed the most clear-cut in terms of item content, thus it was retained in the item set; items 83 and 123 were removed.

Table 2

Anxiety Community Sample Item Parameter Estimates (18 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
3	ANX-A	1.08	0.10	-0.70	0.10	1.43	0.13	3.01	0.26
12	ANX-C	1.24	0.11	0.03	0.08	1.98	0.16	2.94	0.24
30	ANX-P	2.24	0.22	1.20	0.08	1.86	0.12	2.43	0.17
43	ANX-A	1.59	0.14	0.63	0.07	2.02	0.14	2.73	0.21
52	ANX-C	2.48	0.19	0.07	0.05	1.22	0.07	1.85	0.10
70	ANX-P	1.45	0.13	0.46	0.07	1.77	0.13	2.60	0.20
83	ANX-A	0.76	0.08	-1.30	0.17	0.66	0.12	2.65	0.29
92	ANX-C	2.84	0.25	0.65	0.05	1.47	0.08	1.90	0.11
110	ANX-P	0.97	0.09	-1.29	0.14	0.37	0.09	2.25	0.21
123	ANX-A	0.79	0.09	-1.86	0.21	0.22	0.11	2.81	0.30
132	ANX-C	1.93	0.15	0.16	0.06	1.22	0.08	1.86	0.12
150	ANX-P	1.25	0.13	1.04	0.10	2.40	0.21	3.11	0.28
163	ANX-A	1.53	0.13	0.18	0.07	1.48	0.11	2.23	0.16
172	ANX-C	1.78	0.13	-0.74	0.08	0.57	0.06	1.38	0.09
190	ANX-P	0.86	0.10	0.55	0.11	2.39	0.26	3.30	0.36
203	ANX-A	0.94	0.10	-0.22	0.10	1.98	0.19	3.36	0.33
212	ANX-C	1.85	0.19	1.26	0.09	2.01	0.14	2.47	0.19
230	ANX-P	0.57	0.08	-2.30	0.34	0.72	0.16	3.32	0.46

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

Two other item pairs displayed local dependence. For LD between items 3 “I can't do some things well because of nervousness[2]” and 12 “I often have trouble concentrating because I'm nervous[2],” item 12 was retained because of its higher slope (LD χ^2 = 19.4). Interestingly, these items appear on different subscales (item 3 on affective; item 12 on cognitive). The LD between these items indicates that the last half of each item (“because of nervousness”) may be responsible for the LD between these item pairs. Item 132 was removed to resolve the LD between items 132 “My friends say I worry too much[2]” and 172 (LD χ^2 = 14.8). Since both of these items’ slopes were acceptable, item 132 was omitted due to its less direct wording (i.e., this item assumes that one’s friends have commented on one’s mood rather than relying directly on self-report, and friends’ comments may be influenced by any number of other factors).

Next, an item analysis was carried out on the 14 items remaining after the removal of items 3, 83, 123, and 132. This 14-item set was preferable due to the absence of any LD pairs, however, several items’ slope (*a*) parameters were concerning (Table 3). Four of the remaining ANX items had a slope parameter less than one, though two of these items were retained due to the fact that the 95% confidence intervals for the parameters did include one (item 203, 95% CI [0.71, 1.11] and 110, 95% CI [0.73, 1.09]).

Table 3

Anxiety Community Sample Item Parameter Estimates (14 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
12	ANX-C	1.15	0.11	0.03	0.08	2.08	0.18	3.09	0.27
30	ANX-P	2.34	0.24	1.19	0.07	1.84	0.11	2.40	0.17
43	ANX-A	1.65	0.15	0.62	0.07	1.98	0.14	2.68	0.20
52	ANX-C	2.44	0.20	0.07	0.05	1.23	0.07	1.87	0.11
70	ANX-P	1.55	0.14	0.45	0.07	1.71	0.12	2.51	0.19
92	ANX-C	2.73	0.25	0.65	0.05	1.49	0.08	1.93	0.11
110	ANX-P	0.91	0.09	-1.36	0.15	0.39	0.10	2.37	0.23
150	ANX-P	1.29	0.13	1.02	0.10	2.35	0.20	3.04	0.27
163	ANX-A	1.64	0.14	0.18	0.06	1.43	0.10	2.15	0.15
172	ANX-C	1.60	0.12	-0.78	0.08	0.60	0.07	1.45	0.10
190	ANX-P	0.88	0.10	0.54	0.10	2.34	0.25	3.23	0.35
203	ANX-A	0.91	0.10	-0.23	0.10	2.04	0.21	3.46	0.35
212	ANX-C	1.94	0.20	1.24	0.09	1.97	0.14	2.41	0.18
230	ANX-P	0.54	0.08	-2.41	0.37	0.75	0.18	3.48	0.52

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

This led to the need for decisions about items 190 and 230. The very low slope for item 230 led to its removal, but item 190 had a more moderate slope that did not lead to an absolute decision. Due to the absence of any other directly evident problems and the only marginally low slope, item 190 was retained. The deletion of item 230, then, resulted in a potential final item set of 13 items (items 3, 83, 123, 132, and 230 removed). Item parameters are listed in Table 4. An analysis of these 13 items produced a model that fit well (M_2 (689) = 1000.34, $p < .001$; RMSEA = .03).

Table 4

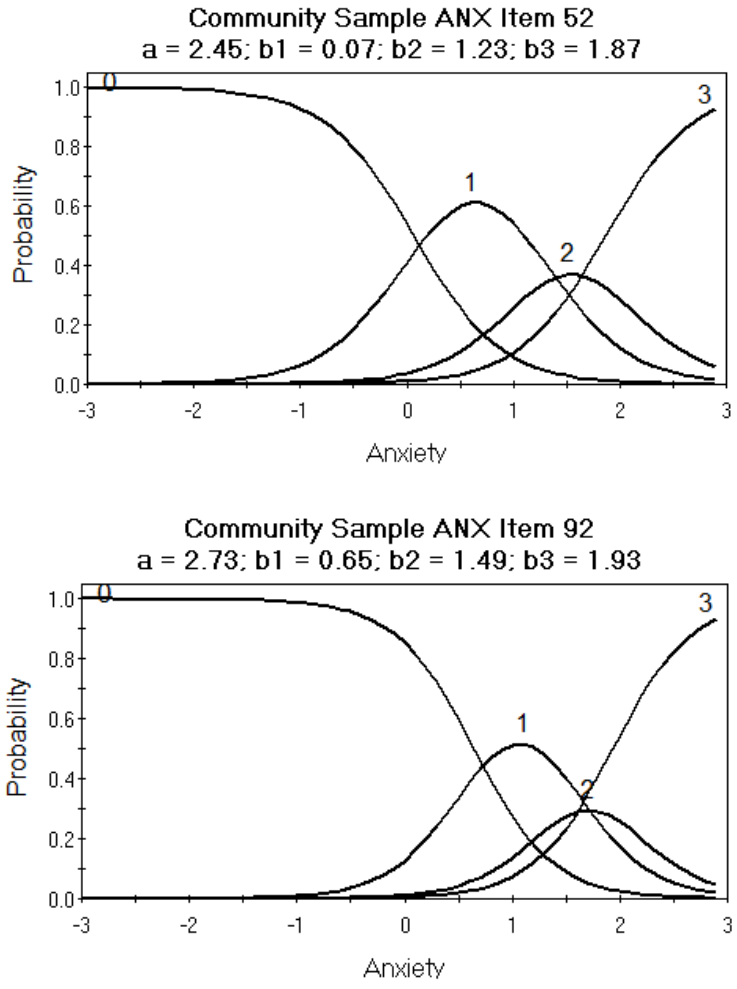
Anxiety Community Sample Item Parameter Estimates (13 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
12	ANX-C	1.14	0.11	0.03	0.08	2.10	0.18	3.12	0.28
30	ANX-P	2.33	0.24	1.19	0.08	1.84	0.12	2.41	0.17
43	ANX-A	1.67	0.15	0.62	0.07	1.97	0.14	2.66	0.20
52	ANX-C	2.45	0.20	0.07	0.05	1.23	0.07	1.87	0.11
70	ANX-P	1.55	0.14	0.45	0.07	1.70	0.13	2.50	0.19
92	ANX-C	2.73	0.25	0.65	0.06	1.49	0.08	1.93	0.11
110	ANX-P	0.88	0.09	-1.39	0.16	0.40	0.10	2.43	0.24
150	ANX-P	1.29	0.13	1.02	0.10	2.35	0.20	3.04	0.27
163	ANX-A	1.65	0.14	0.18	0.06	1.43	0.10	2.14	0.15
172	ANX-C	1.61	0.13	-0.78	0.08	0.60	0.07	1.45	0.10
190	ANX-P	0.87	0.10	0.55	0.11	2.37	0.26	3.27	0.36
203	ANX-A	0.91	0.10	-0.23	0.10	2.04	0.21	3.46	0.35
212	ANX-C	1.93	0.20	1.24	0.09	1.97	0.14	2.42	0.18

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

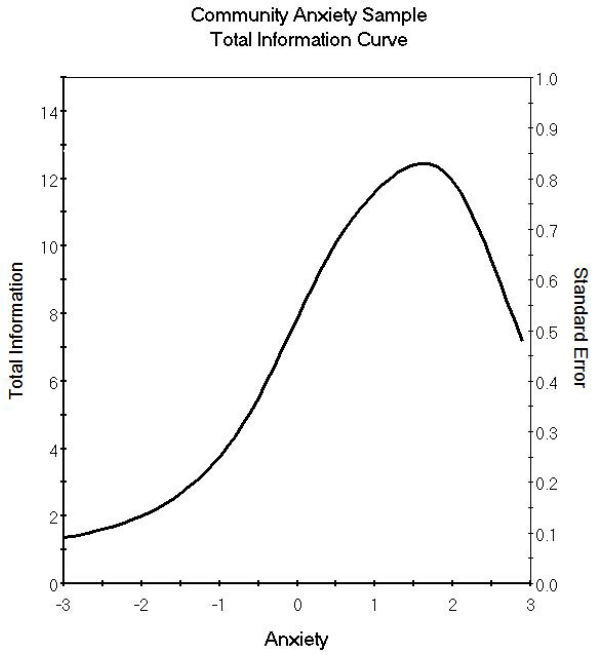
In the community ANX 13-item sample, the two items with the highest slopes are item 52 and item 92. The trace lines for these items are presented in Figure 1. These two items both have item content related to "worry," and their high slope parameters are not surprising since worry seems conceptually related to anxiety. These items demonstrate a relatively stronger relationship between an individual's response to the item and the anxiety continuum, and can be considered more highly discriminating. These two items share similar threshold parameters, or are endorsed at comparable levels of anxiety.

Figure 1. *Trace lines for the Anxiety community sample items with the highest two slopes.*



The test information curve was examined to assess at what point along the continuum of anxiety the revised ANX scale functioned best. Information, in IRT, is directly related to the standard error of measurement for any given level of the underlying construct, and is thus related to the accuracy of an estimate of the underlying construct. A test information curve models the information at each level of the underlying construct, where the peak of the curve indicates the highest information (and lowest standard error). The Anxiety community sample test information curve for the 13-item set appears in Figure 2, and illustrates that the ANX scale yields the greatest information from approximately 0 to +3.0 on the anxiety continuum, and within this range demonstrates a rather large level of total information (approximately 7-12).

Figure 2. *Anxiety test information curve for the community sample.*



Depression

An IRT analysis of all 18 DEP items was conducted on the community sample using the GRM. Item parameters for this full item set appear in Table 5. Problems in the model fit were apparent in the standardized LD χ^2 statistics, though these instances of LD did not appear to be specifically aligned with the subscale structure. Four item pairs indicated problems of local dependence, most notably an LD χ^2 = 68.8 between item 112 and 72. For this pair, content related to sleep seems likely to have contributed to the LD; item 72 was removed due to a slope of less than 1.0. The removal of item 175 resolved the other three of these instances of local dependence. These were with items 205 (LD χ^2 = 14.8), 215 (LD χ^2 = 24.7), and 232 (LD χ^2 = 10.7).

Table 5

Depression Community Sample Item Parameter Estimates (18 Items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	2.01	0.17	0.66	0.06	1.74	0.11	2.47	0.17
15	DEP-C	1.62	0.14	0.31	0.06	1.70	0.12	2.35	0.17
32	DEP-P	1.34	0.14	1.07	0.10	2.48	0.21	3.63	0.35
45	DEP-A	3.40	0.34	1.06	0.06	1.82	0.09	2.48	0.15
55	DEP-C	2.56	0.22	0.79	0.06	1.81	0.10	2.36	0.15
72	DEP-P	0.91	0.09	-1.10	0.14	0.26	0.09	1.88	0.19
85	DEP-A	1.93	0.17	0.84	0.07	2.00	0.13	2.64	0.19
95	DEP-C	0.99	0.10	-0.41	0.10	1.63	0.16	2.98	0.28
112	DEP-P	1.03	0.10	-0.87	0.11	0.35	0.09	1.55	0.15
125	DEP-A	1.27	0.11	0.27	0.07	1.73	0.14	2.46	0.20
135	DEP-C	2.34	0.20	0.56	0.06	1.87	0.11	2.37	0.15
152	DEP-P	1.81	0.17	0.97	0.08	1.97	0.13	2.62	0.19
165	DEP-A	2.57	0.27	1.42	0.08	2.27	0.15	2.56	0.18
175	DEP-C	1.04	0.10	-0.40	0.09	1.32	0.13	3.22	0.29
192	DEP-P	0.92	0.10	0.19	0.09	1.85	0.19	2.73	0.27
205	DEP-A	1.43	0.11	-0.82	0.09	0.62	0.07	1.87	0.14
215	DEP-C	0.89	0.09	-0.99	0.13	0.79	0.11	2.67	0.26
232	DEP-P	0.66	0.09	-0.40	0.13	2.17	0.28	4.48	0.58

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

An analysis was conducted on the remaining 16 items (that is, with items 72 and 175 removed). This 16-item analysis did not indicate any problems in the standardized LD χ^2 statistics, though several items were removed from the scale due to their low (a < 1.0) slopes (Table 6). Initially, item 215 as well as item 232 was removed. Two additional items had a slope less than one in the 16-item analysis, but were retained because the 95% confidence interval for these slopes contained one (items 112, 95% CI [0.73, 1.09] and 192, 95% CI [0.72, 1.07]). The resulting 14-item set was analyzed using a GRM; item 112 was subsequently removed due to low slope (a = 0.86, SE = 0.09). At this stage, this final 13-item IRT-derived DEP scale is proposed for Study 3. The model fit was excellent, (M_2 (689) = 985.07, p < .001; RMSEA = .02). Item parameters for this analysis appear in Table 7.

Table 6

Depression Community Sample Item Parameter Estimates (16 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	2.11	0.18	0.65	0.06	1.71	0.10	2.42	0.16
15	DEP-C	1.66	0.14	0.30	0.06	1.68	0.12	2.32	0.16
32	DEP-P	1.40	0.14	1.05	0.10	2.42	0.20	3.53	0.33
45	DEP-A	3.34	0.33	1.06	0.06	1.83	0.10	2.50	0.16
55	DEP-C	2.55	0.22	0.79	0.06	1.82	0.10	2.37	0.15
85	DEP-A	1.99	0.18	0.83	0.07	1.97	0.13	2.61	0.18
95	DEP-C	1.02	0.10	-0.40	0.09	1.60	0.15	2.91	0.27
112	DEP-P	0.91	0.09	-0.96	0.13	0.38	0.10	1.70	0.18
125	DEP-A	1.31	0.12	0.26	0.07	1.70	0.14	2.41	0.19
135	DEP-C	2.40	0.20	0.55	0.06	1.86	0.11	2.35	0.15
152	DEP-P	1.87	0.17	0.95	0.07	1.94	0.13	2.58	0.19
165	DEP-A	2.56	0.28	1.43	0.09	2.27	0.15	2.57	0.18
192	DEP-P	0.92	0.10	0.19	0.09	1.85	0.19	2.74	0.28
205	DEP-A	1.32	0.11	-0.86	0.09	0.65	0.08	1.97	0.15
215	DEP-C	0.80	0.09	-1.07	0.15	0.85	0.13	2.90	0.31
232	DEP-P	0.61	0.08	-0.43	0.14	2.33	0.32	4.81	0.66

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

Table 7

Depression Community Sample Item Parameter Estimates (13 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	2.20	0.19	0.64	0.06	1.69	0.10	2.38	0.15
15	DEP-C	1.67	0.14	0.30	0.06	1.68	0.12	2.33	0.16
32	DEP-P	1.40	0.14	1.05	0.10	2.42	0.20	3.53	0.33
45	DEP-A	3.20	0.32	1.07	0.06	1.86	0.10	2.55	0.16
55	DEP-C	2.54	0.23	0.79	0.06	1.82	0.11	2.38	0.15
85	DEP-A	1.98	0.18	0.83	0.07	1.98	0.13	2.62	0.19
95	DEP-C	1.09	0.10	-0.39	0.09	1.53	0.14	2.78	0.25
125	DEP-A	1.30	0.12	0.26	0.07	1.70	0.14	2.42	0.20
135	DEP-C	2.40	0.21	0.55	0.06	1.86	0.11	2.36	0.15
152	DEP-P	1.90	0.17	0.94	0.07	1.93	0.13	2.56	0.18
165	DEP-A	2.48	0.27	1.44	0.09	2.31	0.16	2.61	0.19
192	DEP-P	0.88	0.10	0.19	0.10	1.92	0.20	2.84	0.30
205	DEP-A	1.18	0.10	-0.92	0.10	0.69	0.09	2.12	0.17

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

In the community DEP 13-item sample, items 45 and 55 demonstrate the highest slope parameters (Figure 3). In this community sample, the trace lines for item 45 (Figure 3, top panel) are especially well-defined, showing strong distinction among response options. The test information curve for the 13-item set (Figure 4) shows higher information for depression levels in the +0.5 to +3.0 range, yielding information of approximately 8 to 16.

Figure 3. *Trace lines for the Depression community sample items with the highest two slopes.*

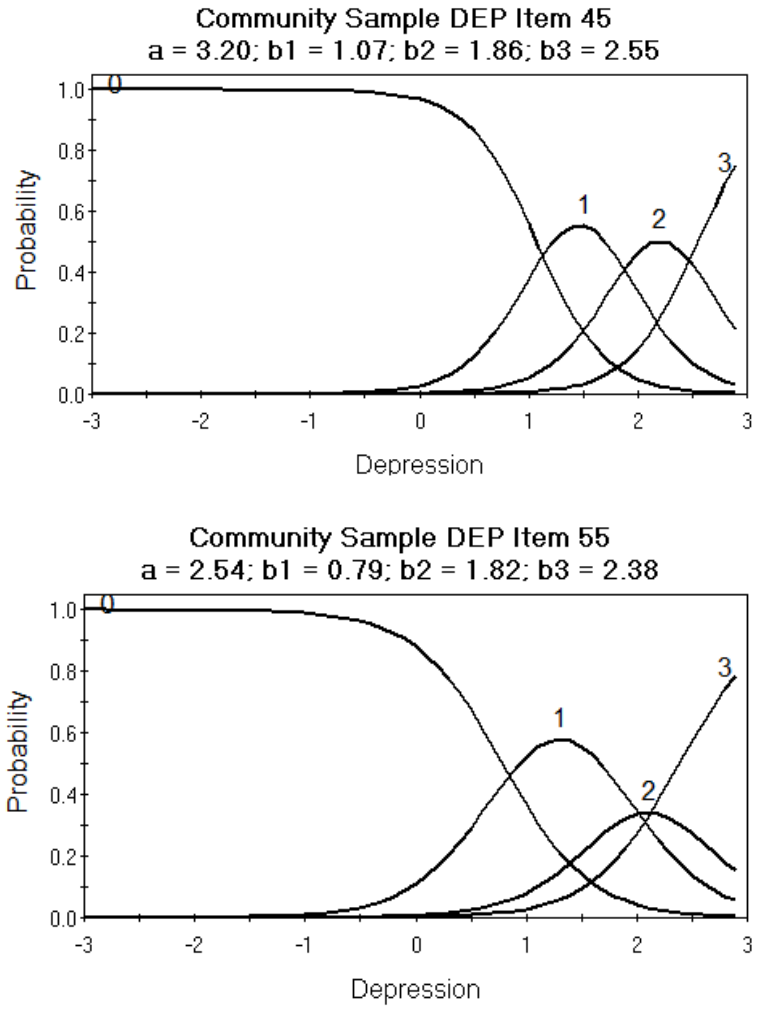
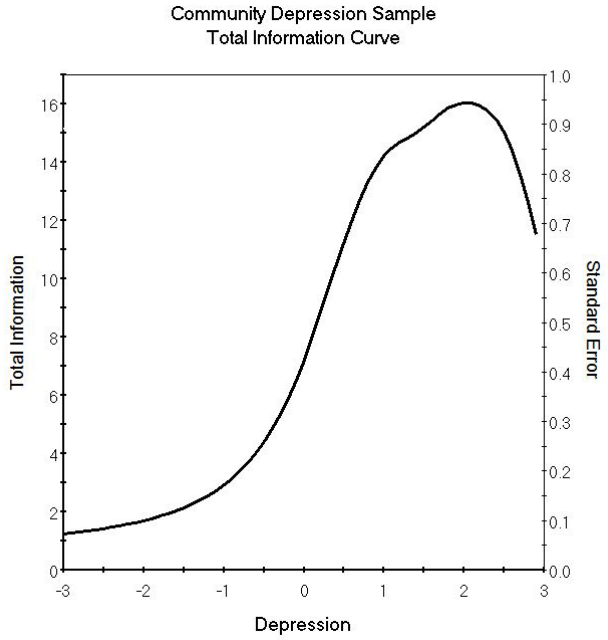


Figure 4. *Depression test information curve for the community sample*.



Study 2: Clinical Sample

PAI-A Clinical Sample

The PAI-A clinical sample of adolescents is a subset of the data collected for the standardization of the PAI-A, used with permission. The PAI-A clinical sample of adolescents (N = 1160, 58.4% male) was obtained from over 75 clinical locations, including sites in outpatient mental health or medical settings, inpatient mental health or medical settings, correctional settings, and school counseling settings. Demographics of this sample appear in Table 8.

Table 8

Demographics of the PAI-A Clinical Sample of Adolescents

Age			Ethnicity		
Age	*n*	Percent	Ethnicity	*n*	Percent
12	45	3.9	White	826	71.2
13	133	11.5	African-American	226	19.5
14	181	15.6	Hispanic	52	4.5
15	240	20.7	Asian	11	0.9
16	237	20.4	Native American	9	0.8
17	213	18.4	Other	19	1.6
18	181	7.0			
Missing	30	2.6	Missing	17	1.5

Results

Anxiety

Using response data obtained from the clinical sample, the 18 items of the ANX scale were also examined using a GRM. The item parameters for the 18-item analysis appear in Table 9. To remove instances of LD in the 18-item analysis, several items were eliminated. As in the community sample, there was LD between items 110 and 123 (LD $\chi^2 = 62.9$) that suggested the removal of item 123 (due to a slope less than 1.0). Items 3 “I can't do some things well because of nervousness[2]” and 12 “I often have trouble concentrating because I'm nervous[2]” again formed a LD pair (LD $\chi^2 = 22.3$), with

Table 9

Anxiety Clinical Sample Item Parameter Estimates (18 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
3	ANX-A	1.64	0.10	-0.28	0.05	0.86	0.06	1.61	0.09
12	ANX-C	1.97	0.11	-0.41	0.05	0.77	0.05	1.47	0.07
30	ANX-P	2.33	0.15	0.57	0.04	1.25	0.06	1.70	0.08
43	ANX-A	1.91	0.11	0.30	0.05	1.19	0.06	1.82	0.09
52	ANX-C	2.45	0.13	-0.52	0.05	0.52	0.04	1.18	0.06
70	ANX-P	1.69	0.10	-0.08	0.05	0.85	0.06	1.57	0.08
83	ANX-A	0.64	0.06	-1.52	0.17	0.26	0.10	2.27	0.23
92	ANX-C	2.86	0.17	0.05	0.04	0.77	0.04	1.29	0.06
110	ANX-P	0.82	0.07	-1.86	0.16	-0.48	0.09	1.48	0.13
123	ANX-A	0.71	0.06	-2.81	0.26	-0.65	0.11	1.58	0.16
132	ANX-C	1.69	0.10	-0.02	0.05	0.86	0.06	1.57	0.08
150	ANX-P	1.74	0.11	0.40	0.05	1.26	0.07	1.85	0.10
163	ANX-A	1.76	0.10	-0.35	0.05	0.64	0.05	1.34	0.07
172	ANX-C	1.78	0.10	-0.84	0.06	0.25	0.05	1.02	0.06
190	ANX-P	0.92	0.07	-0.15	0.07	1.14	0.10	2.18	0.17
203	ANX-A	1.18	0.08	-0.14	0.06	1.34	0.09	2.29	0.15
212	ANX-C	1.55	0.11	0.94	0.06	1.67	0.10	2.28	0.13
230	ANX-P	0.53	0.06	-2.69	0.32	-0.17	0.12	2.33	0.27

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

item 3 removed for its lower slope. Items 132 “My friends say I worry too much[2]” and 172 formed a LD pair (LD χ^2 = 17.4) where item 132 was removed for its less direct wording. Though the item pairs are different than those in the community sample, item 83 was again removed for LD. Item 83 exhibited LD with items 43 (LD χ^2 = 11.9) and 230 (LD χ^2 = 10.9). In addition to these concerns, item 230, eliminated in the community sample for low slope, was removed in this sample due to low slope as well as LD with item 110 (LD χ^2 = 14.4).

A GRM was fit to the remaining 13 items with close fit of the model, (M_2 (689) = 1494.90, $p < .001$; RMSEA = .03). No item pairs were locally dependent, though two items had a slope less than one (Table 10). Item 190 was not deleted since its 95% confidence interval includes one, 95% CI [0.79, 1.07]. This reinforced the decision made in the community sample to retain this item. Item 110 had a moderately low slope, however, in order to take a more conservative approach to item removal, it was also retained.

Table 10

Anxiety Clinical Sample Item Parameter Estimates (13 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
12	ANX-C	1.84	0.10	-0.43	0.05	0.79	0.05	1.52	0.08
30	ANX-P	2.40	0.16	0.56	0.04	1.23	0.06	1.68	0.08
43	ANX-A	1.91	0.12	0.29	0.05	1.19	0.06	1.82	0.09
52	ANX-C	2.45	0.14	-0.53	0.05	0.52	0.04	1.18	0.06
70	ANX-P	1.73	0.10	-0.07	0.05	0.85	0.06	1.56	0.08
92	ANX-C	2.74	0.16	0.05	0.04	0.78	0.04	1.31	0.06
110	ANX-P	0.77	0.07	-1.96	0.18	-0.51	0.09	1.56	0.14
150	ANX-P	1.78	0.11	0.40	0.05	1.25	0.07	1.82	0.09
163	ANX-A	1.86	0.11	-0.34	0.05	0.62	0.05	1.31	0.07
172	ANX-C	1.72	0.10	-0.86	0.06	0.25	0.05	1.04	0.06
190	ANX-P	0.93	0.07	-0.15	0.07	1.13	0.10	2.16	0.16
203	ANX-A	1.17	0.08	-0.14	0.06	1.34	0.09	2.31	0.15
212	ANX-C	1.63	0.12	0.92	0.06	1.63	0.09	2.22	0.13

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

In the clinical sample the two items with the highest slopes were identical to those in the community sample: item 52 and item 92. The trace lines for these items are presented in Figure 5. As in the community sample, the trace lines for these two items are well-defined, or show greater distinction between response categories. The clinical sample test information curve for the ANX 13-item set appears in Figure 6. In this sample, ANX yields information equal to roughly 7-14 from around -0.5 to +2.5 on the anxiety continuum.

Figure 5. *Trace lines for the Anxiety clinical sample items with the highest two slopes.*

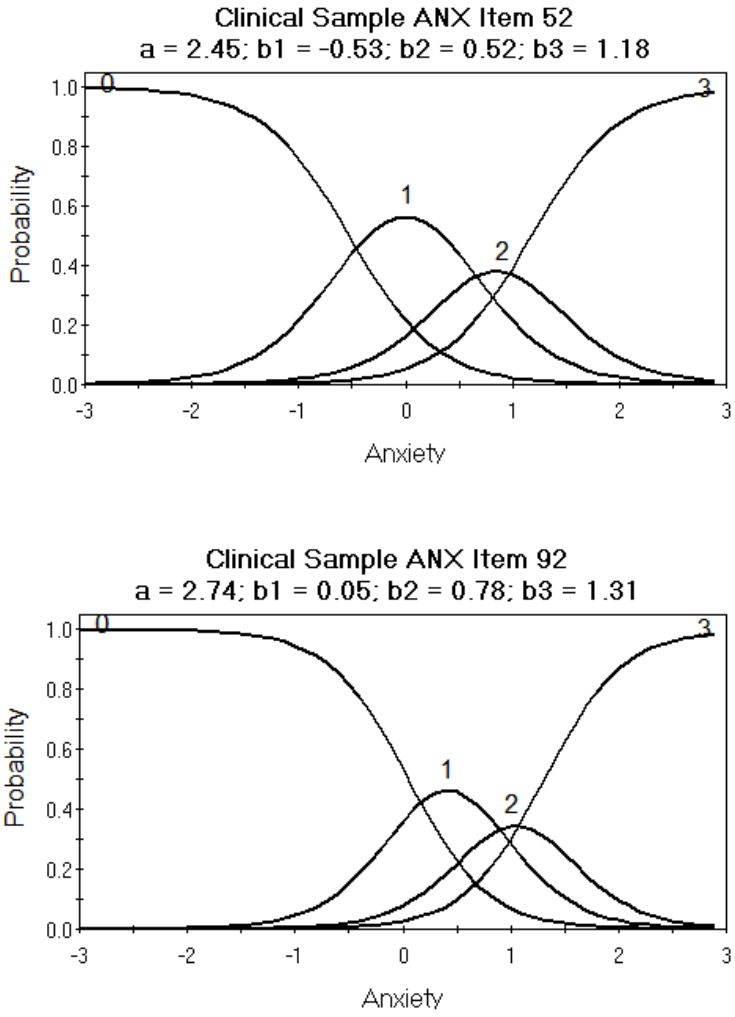
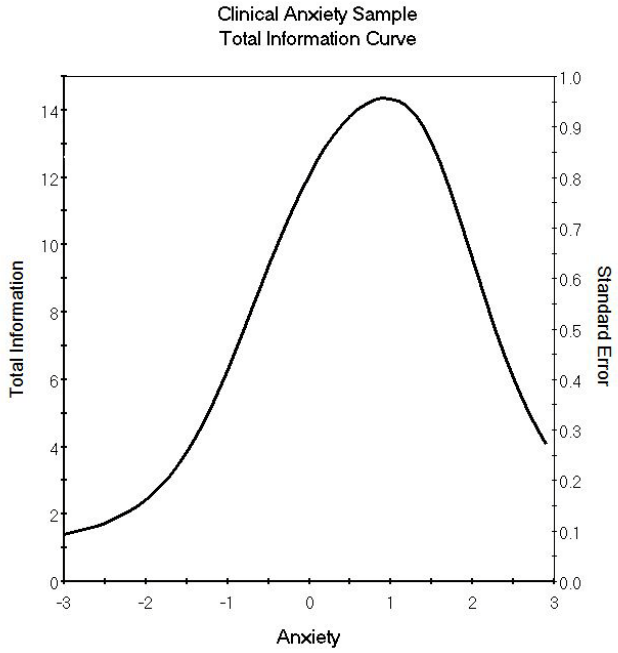


Figure 6. *Anxiety test information curve for the clinical sample.*



Depression

GRM parameters from an analysis of all 18 items appear in Table 11. Several LD pairs suggested problems in this model. As with the community sample, the removal of item 175 resolved several LD pairs (LD χ^2 = 11.1 with item 205, LD χ^2 = 28.0 with item 215, and LD χ^2 = 13.3 with item 232). Ultimately, item 232 was removed based on low slope and to resolve another LD pair (LD χ^2 = 11.8 with item 215). The remaining LD pair (LD χ^2 =96.2) was between items 72 and 112. Of these two items, item 72 was removed due to its lower slope.

Table 11

Depression Clinical Sample Item Parameter Estimates (18 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	1.77	0.11	0.06	0.05	0.96	0.06	1.63	0.08
15	DEP-C	2.04	0.12	-0.15	0.05	0.76	0.05	1.32	0.07
32	DEP-P	1.45	0.10	0.33	0.05	1.50	0.09	2.28	0.13
45	DEP-A	2.23	0.13	0.33	0.04	1.15	0.06	1.73	0.08
55	DEP-C	2.67	0.16	0.09	0.04	0.87	0.05	1.37	0.06
72	DEP-P	0.75	0.07	-1.79	0.17	-0.50	0.09	1.00	0.12
85	DEP-A	2.19	0.13	0.12	0.04	1.08	0.05	1.80	0.08
95	DEP-C	1.27	0.08	-0.81	0.07	0.57	0.06	1.48	0.09
112	DEP-P	0.81	0.07	-1.76	0.16	-0.69	0.10	0.66	0.09
125	DEP-A	1.42	0.09	-0.62	0.06	0.50	0.06	1.24	0.08
135	DEP-C	2.34	0.13	0.02	0.04	0.93	0.05	1.55	0.07
152	DEP-P	1.36	0.09	0.40	0.06	1.42	0.09	2.22	0.13
165	DEP-A	2.40	0.15	0.65	0.04	1.38	0.06	1.91	0.09
175	DEP-C	0.85	0.07	-0.67	0.09	0.82	0.09	2.39	0.19
192	DEP-P	0.86	0.07	-0.42	0.08	0.96	0.10	1.77	0.15
205	DEP-A	1.21	0.08	-1.85	0.12	-0.42	0.06	0.83	0.07
215	DEP-C	0.53	0.06	-2.61	0.31	-0.37	0.12	2.44	0.29
232	DEP-P	0.63	0.06	-0.72	0.12	1.19	0.14	3.19	0.32

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

A GRM using the resulting 15-item DEP scale showed one additional LD pair, between items 215 and 205 (LD χ^2 = 10.7). Item 215 was removed due to a slope of less than one (a = 0.47, SE = 0.06), and a GRM was fit to the remaining 14 items. There was no evidence of local dependence in this analysis, though two slopes were concerning (Table 12). Item 112 was removed for its low slope, leaving 13 DEP items. Item 192, however, was retained due to a moderate slope. This 13-item model fit well, (M_2 (689) = 1244.77, p < .001; RMSEA = .03); parameters appear in Table 13.

For the clinical sample, the DEP items with the two highest slope parameters were item 55 and 135 (Figure 7). These items were highly discriminating, and demonstrated strong relationships with depression. The test information curve for the 13-item set (Figure 8) shows greater information in the higher range of depression. In the clinical sample, information is roughly 7 to 14 between -0.5 to +2.5 on the construct continuum.

Table 12

Depression Clinical Sample Item Parameter Estimates (14 items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	1.80	0.11	0.06	0.05	0.95	0.06	1.62	0.08
15	DEP-C	2.09	0.12	-0.15	0.05	0.75	0.05	1.31	0.07
32	DEP-P	1.44	0.10	0.33	0.05	1.50	0.09	2.28	0.13
45	DEP-A	2.25	0.14	0.33	0.04	1.14	0.06	1.72	0.08
55	DEP-C	2.66	0.16	0.09	0.04	0.87	0.05	1.38	0.06
85	DEP-A	2.21	0.13	0.12	0.04	1.08	0.06	1.79	0.08
95	DEP-C	1.30	0.08	-0.80	0.07	0.56	0.06	1.46	0.09
112	DEP-P	0.72	0.07	-1.95	0.19	-0.77	0.11	0.73	0.11
125	DEP-A	1.46	0.09	-0.61	0.06	0.49	0.06	1.22	0.08
135	DEP-C	2.33	0.13	0.02	0.04	0.93	0.05	1.56	0.07
152	DEP-P	1.38	0.09	0.40	0.06	1.40	0.09	2.19	0.13
165	DEP-A	2.31	0.15	0.65	0.05	1.40	0.07	1.95	0.09
192	DEP-P	0.86	0.07	-0.42	0.08	0.95	0.10	1.76	0.15
205	DEP-A	1.12	0.08	-1.93	0.13	-0.44	0.07	0.87	0.08

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

Table 13

Depression Clinical Sample Item Parameter Estimates (13 Items)

Item	Subscale	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*
5	DEP-A	1.79	0.11	0.06	0.05	0.95	0.06	1.63	0.08
15	DEP-C	2.09	0.12	-0.15	0.05	0.75	0.05	1.31	0.07
32	DEP-P	1.44	0.10	0.33	0.05	1.50	0.09	2.29	0.13
45	DEP-A	2.25	0.14	0.33	0.04	1.14	0.06	1.72	0.08
55	DEP-C	2.69	0.16	0.09	0.04	0.87	0.05	1.38	0.06
85	DEP-A	2.21	0.13	0.12	0.04	1.08	0.06	1.80	0.08
95	DEP-C	1.30	0.08	-0.80	0.07	0.56	0.06	1.46	0.09
125	DEP-A	1.47	0.09	-0.61	0.06	0.49	0.06	1.22	0.08
135	DEP-C	2.34	0.13	0.02	0.04	0.93	0.05	1.56	0.07
152	DEP-P	1.39	0.09	0.40	0.06	1.40	0.09	2.19	0.13
165	DEP-A	2.29	0.15	0.66	0.05	1.41	0.07	1.95	0.09
192	DEP-P	0.85	0.07	-0.43	0.09	0.97	0.10	1.79	0.15
205	DEP-A	1.11	0.08	-1.95	0.14	-0.44	0.07	0.88	0.08

Note. a = slope parameter; b_i = threshold parameters; *s.e.* = standard error

Figure 7. *Trace lines for the Depression clinical sample items with the highest two slopes.*

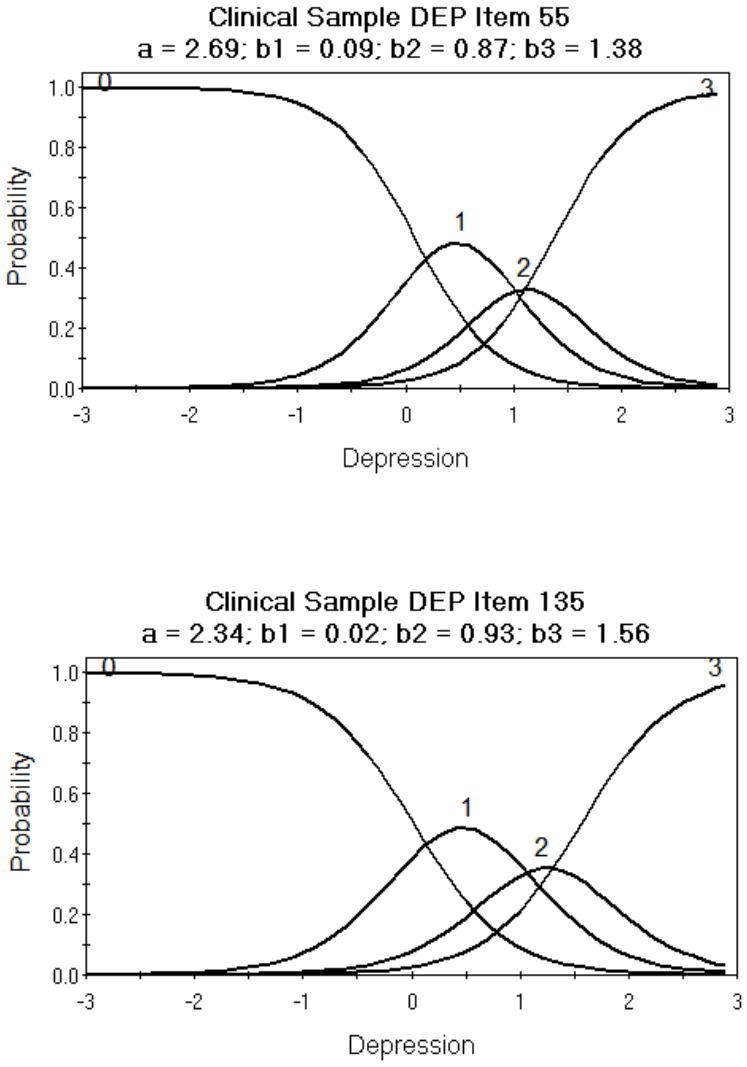
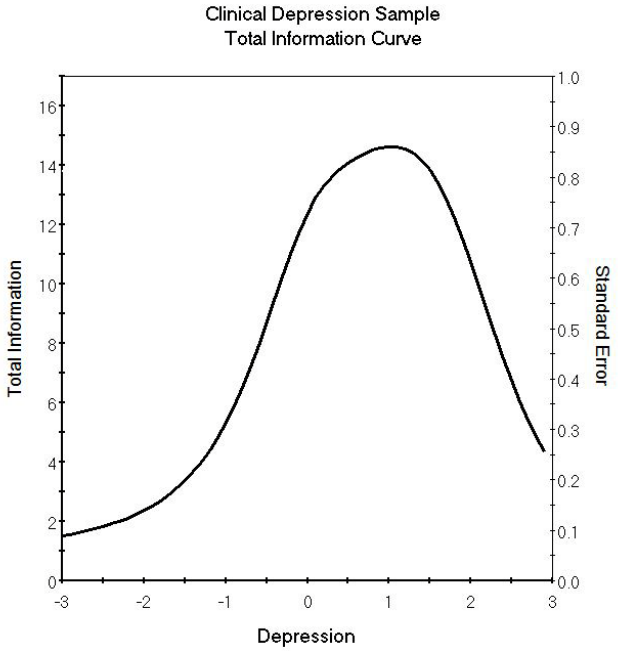


Figure 8. *Depression test information curve for the clinical sample.*



Summary of Study 1 and 2

For both ANX and DEP, the analyses of both the community and clinical samples resulted in the same 13-item set, though with some differences in the logic for item deletion between the two samples. The 13 items that compose the final IRT-derived ANX scale will be used in Study 3 (items 3, 83, 123, 132, and 230 removed). Items removed came from all three subscales: one item each from the cognitive and physiological subscales, and three items from the affective subscale. This seems reasonable; even while there was not strong evidence for the subscales in the 18 item analysis, deleting all items in a single subscale might be unwise. For DEP, the 13-item final scale omits items 72, 112, 175, 215, and 232. In the full length PAI-A, these items fall into the cognitive subscale (2 items) and the physiological subscale (3 items). All items on the affective

subscale were retained.

The items' location parameters place them near the upper end of severity, indicating that high levels of anxiety or depression are needed to endorse the items. Thus, both the clinical and the community samples provide information in the more severe ranges of the constructs. It is typical for most scales to provide better information in a band of the underlying construct, and clinically-purposed scales in particular seem often to provide more information in the severe ranges of a construct (Reise & Waller, 2009).

The test characteristic curve, or expected score curve, graphs the predicted score against the latent trait. In situations where it is desirable to use the traditional summed scoring, our interest lies in the linear relationship between the summed scores and the IRT-scaled scores. When this curve departs significantly from linearity, the traditional summed score may not be a good approximation for the scale-score. Because Study 3 proposes to use the traditional summed score, it is important that this is an adequate score approximation. For ANX and DEP, both the community and clinical samples are linear in the higher ranges of the construct, or the areas where the scale yields more information (Figures 9 and 10), thus summed scores will be used in Study 3.

Figure 9. *Expected score curves for the community samples.*

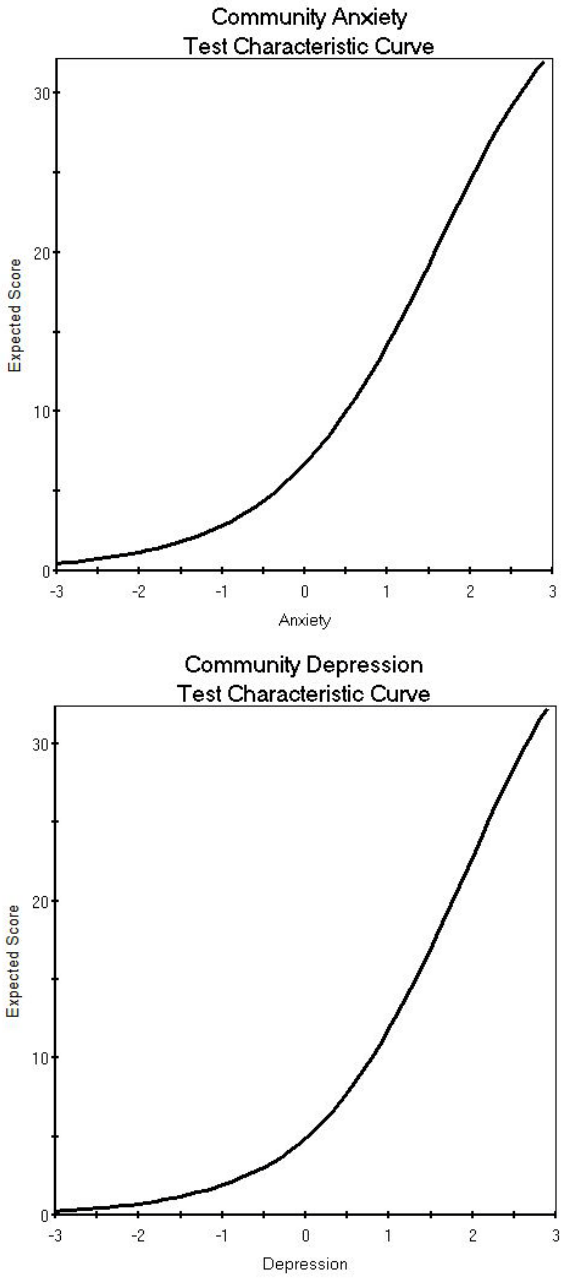
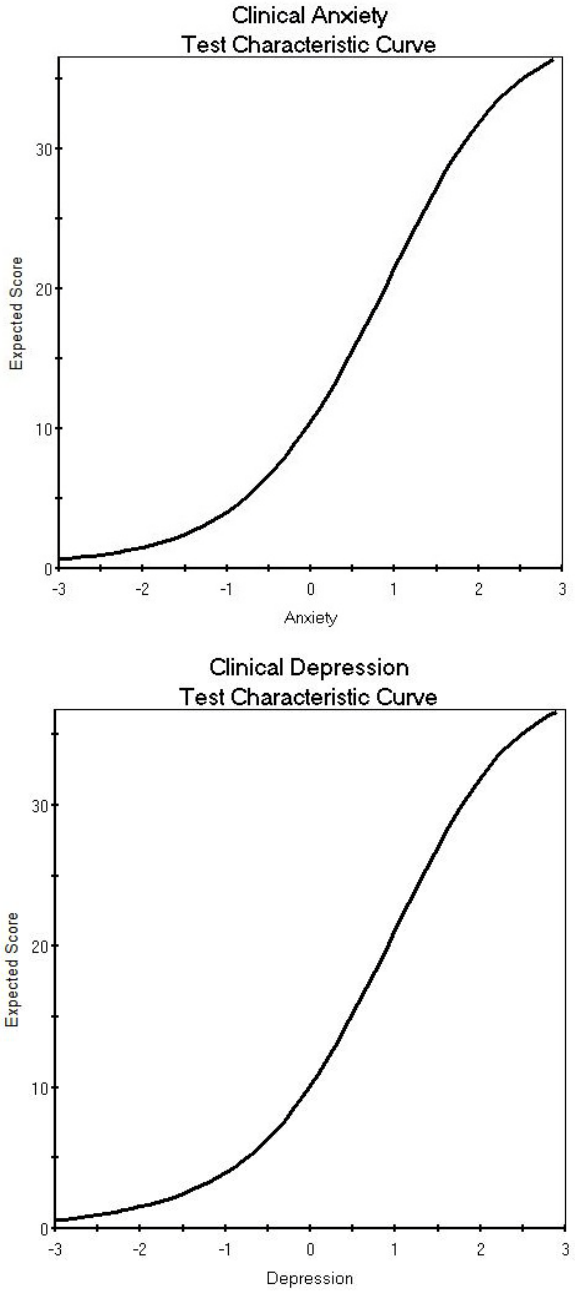


Figure 10. *Expected score curves for the clinical samples.*



Study 3: Construct Validation of the Revised ANX and DEP scales

The evidence for the construct validity (Messick, 1995) of the IRT-derived ANX and DEP scales was examined using data collected from adolescents admitted to an inpatient treatment facility. The two IRT-derived PAI-A scales were expected to exhibit relationships with validating instruments similar to those of the full-length PAI-A scales (i.e., to measure anxiety or depression as well as other instruments that suitably measure these constructs). To assess these relationships in continuous measures, correlations were computed between the validating instruments and the corresponding full-length PAI-A scale. These correlations were repeated for the IRT-derived ANX and DEP scales. Construct validity was assessed by comparing correlations with each validating instrument for the full-length PAI-A scales and the IRT-derived scales. Where available, correlations reported in previous literature were expected to be replicated in the correlations obtained, both for the full-length and IRT-derived scales. For example, the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996) was expected to correlate well with the full-length PAI-A DEP scale; in the revised scale, this correlation should be similar. In categorical measures (i.e., Y-DISC and P-DISC), *t*-tests were conducted to compare the mean differences between 2 categories; specifically "no diagnosis" or "intermediate diagnosis" versus "positive diagnosis." Construct validity was assessed by comparing the *t*-statistic for the original length and IRT-revised PAI-A, where significant values of the *t*-statistic for the original length PAI-A should be reflected in the IRT-revised PAI-A.

There are few studies examining correlations between the PAI-A and the other instruments used for construct validation; all of these use the Beck Depression Inventory

(BDI) or Beck Depression Inventory–II (BDI-II) as the correlate. Additionally, most of the available studies use the PAI rather than the PAI-A, so a different (i.e., longer) item set was used. The correlations reported in the PAI-A professional manual (Morey, 2007) use the BDI, where DEP correlates quite well (*r* =.80). Additionally, several studies report correlations between the adult PAI DEP scale and the BDI or BDI-II. In a sample of chronic pain patients, the PAI DEP scale correlated well with the BDI-II (*r* = .71); ANX also displayed a strong correlation (*r* = .64) with the BDI-II (Hopwood, Creech, et al., 2007). In outpatient samples, Mogge and colleagues (2008) correlated the PAI DEP with the BDI-II (*r* = .86), and Romain (2001) correlated the PAI DEP with the BDI (*r* = .80).

For the present study, comparisons with existing measures included the assessment of correlations of the DEP scale scores with the BDI-II, and the Affective Problems subscale of the Child Behavior Checklist (CBCL/6-18) and Youth Self-Report (YSR; Achenbach, 2007; Achenbach & Rescorla, 2001). Comparisons using *t*-statistics were made for the Major Depressive Disorder subscale of the youth and parent versions of the Diagnostic Interview Schedule for Children (DISC; Shaffer, Fisher, & Lucas, 2004; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000). For ANX, the Multidimensional Anxiety Scale for Children (MASC; March & Parker, 2004), and the Anxiety Problems subscales of the CBCL/6-18 and YSR were all expected to exhibit positive correlations. The categorical youth and parent versions of the DISC (Generalized Anxiety Disorder subscale score) were examined using *t*-tests.

The instruments used for construct validation cover several measurement types: both youth-report (BDI-II, MASC, Youth DISC, and YSR) and parent-report (Parent

DISC and CBCL/6-18) measures, structured interview (both DISC forms) and self-report measures (MASC, BDI-II, CBCL/6-18, and YSR), and continuously (MASC, BDI-II, CBCL/6-18, and YSR), and categorically (both DISC forms) scored measures.

Method

Construct Validation Sample

The construct validation sample was collected adolescents admitted to the adolescent program of a private tertiary care inpatient treatment facility (N = 199). Patients who gave consent were included if they were between 12 and 17 years old and spoke English as their first language. Patients were excluded if they displayed active psychosis, had an IQ less than 70, or were diagnosed with an autism spectrum disorder. Related to the study, patients were administered (among others) the PAI-A, Diagnostic Interview Schedule for Children-Youth Form, Youth Self-Report, Multidimensional Anxiety Scale for Children, and the Beck Depression Inventory-II. A parent/guardian completed the Child Behavior Checklist and the Diagnostic Interview Schedule for Children-Parent Form. Interviewers and other clinicians were blind to each adolescent's diagnoses during the assessment phase.

Measures

Diagnostic Interview Schedule for Children. The Diagnostic Interview Schedule for Children (DISC) is a highly structured interview with both a youth form (Y-DISC; age 9 – 17) and a parent form (P-DISC; for youths age 6 – 17). The interview consists of six sections: anxiety disorders, mood disorders, disruptive disorders, substance-use disorders, schizophrenia, and miscellaneous disorders; a global "whole life" section follows. The DISC (youth form) contains a possible 2930 questions; a minimum

of 358 are asked of each respondent, with additional questions that "stem" from these where applicable. The interview asks about symptoms in the past 12 months and in the past 4 weeks, and the majority of questions are answered with yes/no responses. One-year inter-rater diagnostic agreement for the parent or youth form is acceptable (ranging from $\kappa = 0.65 – 0.92$) for anxiety and depression (Shaffer, et al., 2004; Shaffer, et al., 2000).

The DISC (both youth and parent forms) are largely diagnosis- or categorically-focused, so categorical variables were selected over continuous variables for the DISC forms. Decisions were also made between more general variables of the DISC (e.g., generalized anxiety disorder, major depression) and specific variables (e.g., specific phobia, post-traumatic stress, separation anxiety, dysthymia). Because the PAI-A ANX and DEP scales most closely map onto the diagnostic constructs of generalized anxiety disorder and major depression, these two variables were selected for construct validation using the DISC. These DISC variables are, however, recorded in three categories (negative diagnosis, intermediate diagnosis, positive diagnosis), and were transformed into dichotomous variables such that "negative diagnosis" and "intermediate diagnosis" were collapsed into one group. This method of dichotomizing the variables seemed preferable due to the clinical nature of the construct validation sample; the positive diagnosis categories themselves have sufficient frequencies in most cases.

Child Behavior Checklist. The Child Behavior Checklist (CBCL/6-18), part of the Achenbach System of Empirically Based Assessment, is a parent-report instrument that is a combination of a 120-item checklist for emotional/behavioral difficulties and a 20-item social competency checklist. Subscales are either empirically or DSM-IV-

diagnosis based. The Anxious/Depressed (α = .84) and Withdrawn/Depressed (α = .80) empirically-based subscales are designed to measure anxiety and depression, respectively. Two DSM-IV-based scales also measure anxiety and depression. Anxiety Problems (α = .72) measures symptomology associated with generalized anxiety disorder, separation anxiety disorder, and specific phobias; the Affective Problems subscale (α = .82) is designed to measure symptoms of dysthymia and major depressive disorder (Achenbach, 2007; Achenbach & Rescorla, 2001; Greenbaum, Dedrick, & Lipien, 2004). Scores on the Affective Problems and Anxiety Problems DSM-IV-based scales of the CBCL/6-18 have been shown to predict inclusion in clinically defined depression and anxiety groups (Lengua, Sadowski, Friedrich, & Fisher, 2001).

Youth Self-Report. The Youth Self-Report (YSR), also part of the Achenbach System of Empirically Based Assessment, is a self-report measure for use with adolescents from ages 12 to 18. The YSR measures areas of general competence (20 items) and psychopathology (112 items arranged in 8 empirically-based and 6 DSM-IV-based subscales, comparable to the CBCL/6-18). The empirically-based Anxious/Depressed (α = .84) and Withdrawn/Depressed (α = .71) subscales are designed to measure anxiety and depression. The DSM-IV-based Anxiety Problems subscale (α = .67) is designed to assess generalized anxiety disorder, separation anxiety disorder, and specific phobias. The Affective Problems subscale (α = .81) is designed to evaluate symptoms of dysthymia and major depressive disorder (Achenbach, 2007; Achenbach & Rescorla, 2001).

Within the CBCL/6-18 and YSR, a choice also needed to be made between categorical and continuous variables. Because these two measures are designed to be

relatively continuous (and because the DISC provides an excellent categorical variable), the continuous variables were selected. A further decision was required between the empirically-based and the DSM-IV-based scales of the CBCL/6-18 and YSR. Because the empirical scales are derived directly from the available data and do not have an externally imposed structure, these scales are more general (e.g., "anxious" and "depressed" fall together on one scale). The DSM-IV scales have this external structure imposed, and are more precise due to the DSM-IV's exact definition of what warrants a diagnosis. The DSM-IV scales also seem to more closely approximate the general idea of the individual PAI-A ANX and DEP scales; these DSM-IV scales will thus be used rather than the empirically-based scales of the CBCL/6-18 and YSR.

Multidimensional Anxiety Scale for Children. The Multidimensional Anxiety Scale for Children (MASC) is a 39-item self-report anxiety measure for respondents ages 8 – 19. The MASC produces 13 scores, with four main subscales, an anxiety index, an inconsistency index, and a traditional summed total MASC score ($\alpha = .90$). These four major subscales include: Physical Symptoms, Harm Avoidance, Social Anxiety, and Separation/Panic; the first three of these subscales are further divided into Tense Symptoms and Somatic Symptoms, Perfectionism and Anxious Coping, and Humiliation Fears and Performance Fears. The total summed score (rather than the anxiety index) will be used for the correlations involving the MASC.

Beck Depression Inventory-II. The Beck Depression Inventory-II (BDI-II) is a 21-item self report inventory of depressive symptoms, reported for the most recent two weeks. It is for use with respondents age 13 and over, and is for use in both clinical and nonclinical populations. The BDI-II items ($\alpha = .91$) form one total score, with cut-scores

for “minimal depression” through “severe depression.” The BDI-II is intended for assessing the severity of depressive symptoms rather than the specific presence or absence of depression as a diagnosis (Dozois & Covin, 2004). Scores on the BDI-II correlate well with scores on depression scales of other adolescent instruments, such as the Reynolds Adolescent Depression Scale (Krefetz, Steer, Gulab, & Beck, 2002).

Results

In the construct validation sample, PAI-A profile validity was assessed for all participants (N = 199). This was accomplished using three criteria: the PAI-A’s Inconsistency (ICN) scale, Infrequency (INF) scale, and the frequency of missing responses on the complete PAI-A. ICN is an empirically derived scale designed to reflect consistency in response. It contains 10 item pairs that are usually answered in relation to one another (5 answered in the same direction and 5 answered opposite to one another). ICN is designed to represent “carelessness or confusion” in responses (Morey, 2003, 2007). Scores over 78*t* represent inconsistency in response such that profiles may best be considered invalid, with a completely random completion of the PAI-A resulting in a *t*-score of around 82*t* (Morey, 2007). One participant was excluded based on an ICN *t*-score > 78*t*. INF is another PAI-A validity scale aimed at identifying atypical or idiosyncratic responding. The 7 INF items were designed to be answered in the same direction for all respondents (clinical and nonclinical), some “not true” and some “very true” (Morey, 2003, 2007). These items are designed to be plausible, but either very common or very uncommon. Endorsing these items in an unpredicted direction increasingly points toward lack of attention to the item content, or in some cases, trouble reading in the English language (Morey, 2003). Scores over 79*t* represent inconsistency

in response (Morey, 2007). Two participants were excluded based on a *t*-score > 79*t*. Additionally, Morey (1991, 2003, 2007) recommends that PAI-A profiles that are less than 95% complete be excluded. Six participants were excluded based on PAI-A profiles with responses recorded for fewer than 20% of PAI-A items. The final sample (N = 169) was obtained after omitting participants who declined or revoked consent or who met the exclusion criteria, who did not complete both the PAI-A and at least one other measure, or who did not have a valid PAI-A profile.

Anxiety

Five scales were used in the process of examining evidence for the construct validity of the IRT-derived ANX scale: the MASC total score (N = 122), the youth DISC Generalized Anxiety Disorder scale (N = 165), the parent DISC Generalized Anxiety Disorder scale (N = 160), the YSR Anxiety Problems *t*-score (N = 166), and the CBCL/6-18 Anxiety Problems *t*-score (N = 162). Correlations with each of the continuous scale scores (i.e., MASC, CBCL/6-18, YSR) and both the full-length and IRT-derived PAI-A ANX scale scores appear in Table 14. For the categorical variables (i.e., Y-DISC, P-DISC), a *t*-test was used to test the mean differences between the PAI-A scores in the "no or intermediate diagnosis" and "positive diagnosis" groups.

Table 14

Correlation Coefficients with Validating Anxiety Instruments

	MASC (N = 122)	YSR (N = 166)	CBCL (N = 162)
ANX Full-length	0.73	0.65	0.24
ANX IRT-revised	0.73	0.66	0.24

Note. MASC refers to the Multidimensional Anxiety Scale for Children; YSR refers to the Youth Self-Report Anxiety Problems scale, CBCL refers to the Child Behavior Checklist 6-18 Anxiety Problems scale. All correlations are significant at $p < .01$.

For continuously scored instruments examined using correlations, there were overall differences in terms of relationship with the PAI-A. The MASC displayed the strongest correlation of the measured anxiety scales in both the full-length and IRT-derived scales (r = 0.73). The Anxiety Problems scale of the YSR and of the CBCL/6-18, comparable as self- and parent-report measures, performed quite differently in terms of correlation. For the YSR (self-report) the correlations were strong ($r \approx 0.65$); correlations for the CBCL/6-18 (parent-report) were notably smaller (r = 0.24). The YSR and CBCL/6-18 Anxiety Problems scales correlate only slightly (r = 0.26), suggesting that method variance is contributing to the difference in the correlations between these measures.

In terms of comparison between the full-length and in the IRT-derived scales, the MASC and the CBCL/6-18 demonstrated equally strong correlations between the full-length and the IRT-derived scale. In both the YSR and CBCL/6-18, the full-length and the IRT-derived scale correlations were very similar. The strong relationships

Table 15

Descriptive statistics and independent samples t-tests comparing DISC groups for anxiety

	Negative or Intermediate Diagnosis			Positive Diagnosis					
	N	*M*	*SD*	*N*	*M*	*SD*	*t*	*p*	*d*
Youth-DISC									
ANX Full-length	147	39.22	14.06	18	53.44	12.7	4.09	<.001	1.03
ANX IRT-revised	147	26.90	9.59	18	35.72	8.46	3.72	<.001	0.94
Parent-DISC									
ANX Full-length	138	40.06	14.17	22	47.50	16.93	2.23	.028	0.51
ANX IRT-revised	138	27.54	9.68	22	31.41	11.26	1.70	.091	–

Note. DISC refers to the Diagnostic Interview Schedule for Children; *d* refers to Cohen's *d*; PAI-A responses for each item were scored from 1 to 4.

between the full-length and the IRT-derived scales indicate that the IRT-derived scale is likely a good approximation of the full-length ANX scale.

For the categorical Y-DISC and P-DISC, a *t*-test of mean differences in PAI-A scores significantly distinguished between diagnostic groups (Table 15), where the "positive diagnosis" group scored higher than the "no/intermediate diagnosis" group. The mean differences of the full-length PAI-A scores were significant for both the Y-DISC and P-DISC. The positive diagnosis group had higher PAI-A ANX scores than the no/intermediate diagnosis group. The effect size (Cohen's *d*) for the full-length scales showed a difference between groups of approximately 1 standard deviation for the Y-DISC and of half a standard deviation for the P-DISC. For the IRT-revised version, scores only significantly differentiated diagnostic groups for the Y-DISC; the PAI-A ANX scores did not differentiate between the diagnostic groups in the P-DISC. Cohen's *d* again demonstrated a difference of near 1 standard deviation between groups for the Y-

DISC; no effect size is reported for the P-DISC as the *t*-test was not significant. Kappa (inter-rater reliability) was also calculated for the Y-DISC and P-DISC diagnoses (κ = 0.142). This is a low value, suggesting poor agreement between these two forms of the DISC.

Depression

The five scales used to provide evidence for the validity of the IRT-derived DEP scale included: the BDI-II total score (*N* = 128), the youth DISC Major Depressive Disorder scale (*N* = 165), the parent DISC Major Depressive Disorder scale (*N* = 160), The YSR Affective Problems *t*-score (*N* = 166), and the CBCL/6-18 Affective Problems *t*-score (*N* = 162). Correlations between both the PAI-A DEP full-length and the IRT-derived and the BDI-II, CBCL/6-18, and YSR appear in Table 16.

Table 16

Correlation Coefficients with Validating Depression Instruments

	BDI-II (*N* = 128)	YSR (*N* = 166)	CBCL (*N* = 162)
DEP Full-length	0.84	0.79	0.38
DEP IRT-revised	0.84	0.79	0.37

Note. BDI-II refers to the Beck Depression Inventory-II; YSR refers to the Youth Self-Report Affective Problems scale, CBCL refers to the Child Behavior Checklist 6-18 Affective Problems scale. All correlations are significant at $p < .01$.

For continuously scored measures, relationships between the PAI-A and each instrument varied in magnitude. The BDI-II displayed the strongest correlation with the PAI-A DEP scales (*r* = .84). Previously reported correlations with the BDI (*r* = .80) and BDI-II (*r* = .71 - .86) were largely replicated in this sample. However, none of these

previously reported correlations pair the PAI-A directly with the BDI-II (correlations are reported between the adult PAI and both the BDI and BDI-II, as well as the PAI-A and the BDI), so no direct comparison can be made. The YSR and the CBCL/6-18 correlated differently with the full-length and IRT-derived scales, where the self-report measure (the YSR) correlated better ($r = 0.79$) than the parent-report measure (the CBCL/6-18; $r \approx 0.37$). The *t*-tests for both the Y-DISC and the P-DISC in the original PAI-A were significant at the .05 level, in both the full-length and IRT-derived scale (Table 17). Those with a positive diagnosis had higher scores compared to those with no/intermediate diagnosis. Effect sizes showed a difference between groups of well over 1 standard deviation for both PAI-A forms for the Y-DISC, but of only slightly over a third of a standard deviation for both forms in the P-DISC. The inter-rater reliability between the Y-DISC and P-DISC was slightly higher for DEP ($\kappa = 0.23$), though this value is still quite low, again suggesting method variance as the cause for the differences in the youth- and parent-report forms of the DISC. Thus, for all five instruments used, the original and IRT-derived scales were comparable, suggesting that the IRT-derived DEP scale is a good likeness of the original DEP scale.

Table 17

Descriptive statistics and independent samples t-tests comparing DISC groups for depression

	Negative or Intermediate Diagnosis			Positive Diagnosis					
	N	*M*	*SD*	*N*	*M*	*SD*	*t*	*p*	*d*
Youth-DISC									
DEP Full-length	104	36.35	11.25	58	51.28	9.61	8.52	<.001	1.40
DEP IRT-revised	104	24.62	8.89	58	37.10	7.80	8.95	<.001	1.47
Parent-DISC									
DEP Full-length	82	39.44	11.86	79	44.08	13.55	2.31	.022	0.37
DEP IRT-revised	82	27.46	9.78	79	31.03	10.78	2.20	.030	0.35

Note. DISC refers to the Diagnostic Interview Schedule for Children; *d* refers to Cohen's *d*; PAI-A responses for each item were scored from 1 to 4.

In general, the briefest measures of anxiety (i.e., the MASC) and depression (i.e., the BDI-II) seem to correlate best for both the full-length and IRT-derived DEP scale. These two instruments correlated the highest with the PAI-A original and reduced ANX and DEP scales and displayed identical correlations between these two PAI-A scales. Overall, all ANX and DEP correlations were comparable for the original and IRT-revised scales, even in the validating instruments where the correlations with the PAI-A scales was notably lower (e.g., the CBCL/6-18).

Discussion

The overarching goal of this research was to use IRT to comprehensively evaluate the items of the PAI-A ANX and DEP scales. The study focused on two main issues: (1) using IRT methods to assess the functioning of the ANX and DEP items and to eliminate items that were not functioning well, and (2) providing evidence for the hypothesis that these new item sets perform similarly to the original item sets using a variety of other

anxiety and depression measures. Together, the examination of the PAI-A community standardization and clinical samples (Studies 1 and 2) led to the creation of 13-item sets from the original 18-item ANX and DEP scales. To examine construct validity of these two IRT-derived scales (Study 3), correlations using the full item set and the reduced item set were compared in a separate clinical sample of adolescents. This study provided supporting evidence for the idea that the IRT-revised scales perform as well as the original PAI-A scales.

The initial item pool for the adult PAI (more than 2,200 items) was extensively refined for the publication of the PAI (344 items) and, with a focus on adolescent needs, for the 264-item PAI-A (for a detailed description of the revision process see Morey, 1991, 2007). This refinement used multiple criteria to include items that focus on the full continuum of each construct and that best capture the psychological experiences of the respondent (Morey, 1996, 1999, 2007; Morey & Hopwood, 2006, 2008). Though the PAI-A is a relatively brief measure (considering its 22 nonoverlapping scales), it remains an extensive assessment. Current reduced-length adult PAI administration has been limited to introduction of the PAI-Short Form (PAI-SF) and the Personality Assessment Screener (PAS), both adult measures. While research has evaluated these two promising scales favorably (Creech, et al., 2010; Frazier, Naugle, & Haggerty, 2006; Sinclair et al., 2010; Sinclair et al., 2009), an abbreviated PAI-A remains on the horizon. IRT seems to be an ideal tool for this purpose.

This study extended previous classical test theory analyses of the PAI-A into an item response theory framework for the ANX and DEP scales of the PAI-A. The goal in these IRT analyses was to make sure that all items are contributing to the measurement of

the construct of interest, anxiety or depression. While comprehensive revisions during the initial development of the PAI and PAI-A led to scales made up of exceptional items, IRT models are helpful in identifying any items that might be removed without losing significant scale functioning. Consequently, the analyses conducted in the IRT portion of this research did not have the specific goal of shortening the ANX or DEP scales, but of identifying low functioning in individual items; the recourse to address any problems in item functioning was generally, however, to remove an item.

An IRT analysis of the PAI-A is profitable for several reasons. Most generally, IRT analysis is an item-level process; this contrasts with the largely scale- or test-level focus of classical test theory. Because IRT models response probabilities at the item level, it allows for more precise strategies for the improvement of the specific items in a scale. Also, important to the subscale structure of ANX and DEP, IRT provides a method of assessing the dimensionality of scales. The further investigation of the original and IRT-derived scales in terms of their relationship to other anxiety and depression measures provides evidence regarding the construct validity of the resultant IRT-derived scales.

Major Findings and Conclusions

Studies 1 and 2. The PAI-A community standardization data was used in Study 1 to investigate the IRT graded response models separately for ANX and DEP and to identify items that were not consistent with the rest of the scale. Study 2 repeated these analyses with the PAI-A clinical sample and arrived at the same set of items, though some differences in decision logic were used. Though several item selection criteria were used throughout these analyses, specific attention was given to eliminating LD, removing items with low slope parameters, and verifying final model fit. Additionally, both the

community sample (Study 1) and the clinical sample (Study 2) final item sets needed to be equivalent in order for one item set for ANX and for DEP to progress to Study 3. The community sample was determined to be the most important in decision making due to the large variety of potential diagnoses in the clinical sample, though there was ultimately no conflict between the Study 1 and Study 2 final item-sets. The first concern in the 18-item analyses was determining dimensionality, with attention to the subscales of ANX and DEP (i.e., affective, cognitive, physiological).

The PAI subscales (e.g., affective, cognitive, physiological) were created to more aptly identify separate facets of each clinical scale and to add meaning to elevations on a particular scale (Morey, 1996, 1999). The specific patterns of subscale elevations across the entire PAI-A, especially when considered along with other scale and/or subscale evidence, were designed to help guide in differential diagnosis. However, in the current research, there were no clearly evident subscale-related patterns in the LD for the18-item analyses of ANX and DEP. This lead to the use of unidimensional IRT for all analyses, with no preservation of a subscale structure.

After deciding on a unidimensional IRT model, attention focused on removing LD until a final model with no LD was achieved. There were LD problems in all analyses (for both Study 1 and 2). When looking at any given pair of items, this decision was often made on the basis of low slope in one (or more) item. According to these analyses (i.e., removing LD and items with low slope), each original 18-item scale was reduced to 13 items, though no specific effort was made to make the two IRT-derived scales of equal length.

***Anxiety*.** For both the community and clinical ANX samples, a total of 3 progressive GRM analyses eliminated all significant LD and converged on a final item set, where most eliminated items were removed from the 18-item analysis due to both LD and a low slope. Several other items attracted attention for a slope less than 1.0, though in most cases these items' slopes were only moderately low, with confidence intervals for the slope often containing one. In total, the final 13-item analysis omitted items 3, 83, 123, 132, and 230.

***Depression*.** In the two samples, the items also showed several pairs of LD, which were removed through a set of 4 GRM analyses. These analyses did not progress equally in each sample, with items being removed for different reasons and at different times in each sample. The analyses, however, both resulted in equivalent 13 item sets. The final IRT-derived model fit well, and excluded items 72, 112, 175, 215, and 232.

Study 3-Construct Validity. The reliable and valid assessment of psychopathology in adolescents is crucial, and anxiety and depression specifically are arguably the most prevalent and the most difficult to disentangle. In addition, accurate assessment frequently involves not only diagnosis, sometimes a difficult prospect in itself, but also choices regarding treatment planning (e.g., choice of format, length, or setting), and change or outcome evaluation (Morey, 1999).

While PAI-A scale scores are not intended to correspond directly to DSM-IV-TR diagnoses, the PAI-A as a whole and its scales, subscales, and items were each designed to provide information useful to a clinician and be meaningful in the diagnostic process. The adult PAI has shown specific strengths in its convergent and discriminate validity (Morey, 1991, 1999; Morey & Hopwood, 2008), and the adolescent PAI-A seems to be

proceeding similarly (Morey, 2007). For example, in previous research, the PAI-A ANX and DEP scales displayed convergent validity with scales from the Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A), NEO Five-Factor Inventory (NEO-FFI), Adolescent Psychopathology Scales (APS), Symptom Assesssmnet-45 (SA-45), State-Trait Anxiety Inventory (STAI), and Beck Depression Inventory (BDI), among many others listed by Morey (2007). Still, several common adolescent clinical assessment instruments do not appear to have been investigated in relation to the PAI-A. The current research assessed evidence for the construct validity of ANX and DEP by examining (a) correlations with other continuously scored measures of anxiety and depression and (b) mean differences in ANX and DEP scores depending on the presence or absence of a diagnosis (anxiety or depression), as determined by scores on the youth and parent versions of the DISC. In a clinical sample of adolescents, the ANX and DEP IRT-reduced scales displayed evidence for the hypothesis that the original and IRT-revised scales approximate each other. This was true for comparisons made in both youth- and parent-report measures, structured interview and self-report measures, and continuously and categorically scored measures. While each of these methods has limitations, the use of this multi-method design is preferable to a single method of assessment.

Looking only at the validity of the original, full-length PAI-A, most scales related well to the PAI-A ANX and DEP scores. Correlations in the existing literature are restricted to the BDI and BDI-II, and the BDI-II correlation found in this research was within the range of these previously reported correlations. The MASC and BDI-II performed very well, with particularly high correlations. The youth measure (YSR)

performed better overall than the parent report CBCL/6-18. For the full-length PAI-A scale scores, those with a positive diagnosis on the youth and parent DISC had higher scores compared to those with no/intermediate diagnosis. For the IRT-revised scores, the mean differences were found between the youth DISC categories for ANX and DEP; however, differences between the parent DISC categories were found only for DEP. That is, the positive diagnosis group scored higher than the no/intermediate diagnosis group. These findings lead to several possible conclusions.

The MASC and BDI-II performed best in terms of correlation. Other research has found success with the MASC (March & Parker, 2004; March, Parker, Sullivan, & Stallings, 1997), though one study found discriminant evidence only in females (Dierker, et al., 2001). The BDI-II has also consistently performed well, and remains one of the most popular and successful measures of depression (Kashani, Sherman, Parker, & Reid, 1990; Stulz & Crits-Christoph, 2010).

While other-report (e.g., parent or teacher) is considered particularly important in the assessment of child and adolescent symptomology (De Los Reyes & Kazdin, 2005; Duhig, Renk, Epstein, & Phares, 2000; Kraemer et al., 2003), the scales of the parent-report measures used (P-DISC, CBCL/6-18) did not provide strong evidence for a relationship to the ANX and DEP scales of the PAI-A. Though some difference in youth- and parent- report is expected and even desired to avoid redundancy, it was anticipated that both the youth- and parent-report measures would provide similar evidence for a relationship with the PAI-A ANX and DEP scales (Connelly & Ones, 2010). A meta-analysis by Achenbach, McConaughy and Howell (1987) found differences between self- and other-report measures in children and adolescents, but suggested that these low

correlations might be caused by constructs of interest that are variable across situations rather than by specific problems in instrument reliability or validity. Interestingly, Achenbach et al. (1987) also found that self- and other-report measures in adolescents were less similar than those in children, and that self-other correlations were lower for internalizing disorders than for externalizing disorders (e.g., anxiety and depression). Based on this meta-analysis, the current research (i.e., an investigation of internalizing disorders in adolescent respondents) might, in fact, be somewhat less likely to exhibit a high correlation between the self- and parent-report measures.

The PAI-A scale scores differentiated between "no" or "intermediate" diagnosis and "positive" diagnosis on both DISC forms (structured interview) for both full-length PAI-A scales and in the DEP (but not the ANX) IRT-revised scale. Past research has suggested that structured interviews offer gains over unstructured interviews (Miller, Dasher, Collins, Griffiths, & Brown, 2001), and that various forms of the DISC have been reasonably successful in terms of convergent and discriminant validity (Miller, et al., 2001; Shaffer, et al., 2004; Shaffer, et al., 2000). Other research, however, has shown quite low agreement between DISC youth- and parent-report (see Grills & Ollendick, 2002 for a review), and in the current research the P-DISC did not perform as well as other instruments, specifically for ANX.

In terms of comparison between the IRT-derived and the original PAI-A scale correlations (i.e., MASC, BDI-II, CBCL/6-18, YSR) no scale showed more than minimal variation. For the youth and parent DISC, *t*-tests indicated that the IRT-revised PAI-A worked well, with the exception of the P-DISC DEP scale. This strongly supports the use

of the IRT-derived ANX and DEP scales as measures of anxiety or depression, respectively.

Limitations

There are several important limitations of these three studies. In order to replicate or extend the IRT portion of this study, relatively large samples are needed for accurate parameter estimation. Any replication would require somewhat extensive data collection.

The PAI and PAI-A were designed such that patterns of subscale elevation (e.g., physiological, affective, cognitive) across multiple subscales (sometimes spanning several primary scales, such as ANX and DEP) are useful in arriving at a differential diagnosis (Morey, 1991, 1996, 1999, 2007). While Morey (1996) describes many combinations of subscale elevation in the PAI that can be used to aid in accurate diagnosis, in the PAI-A these subscales are not as apparent. In these samples, there was not strong evidence that the ANX and DEP subscales were truly functioning as distinct subscales, thus decisions regarding item elimination did not consider subscale structure; the subscales in the final item sets were not all of equal length. Because the balance of items was not maintained across subscales, all subscale structure and thus any information gleaned from the subscales is lost in the IRT-derived scales.

Recommendations for Further Research

Some validating instruments, particularly the youth and parent DISC and all parent-report measures, did not display expected correlations with the original or the IRT-derived PAI-A ANX and DEP scales. Additional research using different construct validation samples should investigate these patterns, and determine if they can be replicated.

Additional analyses using differential item functioning (DIF) might have added information about differences at the item level between the community and clinical samples. That is, do any items function differently in the community and in the clinical samples, in terms of their relationship to the construct of interest and as evidenced by parameter differences (Cohen, Kim, & Baker, 1993; Thissen, Steinberg, & Wainer, 1993)? If some number of items did exhibit DIF, they could provide evidence for the removal of additional items, or alternately, for the removal of different items as choices were made in the resolution of LD.

Additional analyses using DIF could lead toward the further refinement of the item sets, and could suggest different item sets. This analysis might be particularly helpful between samples of clinical and nonclinical participants, and between males and females. Future studies may examine additional scales of the PAI-A, both in terms of IRT measurement models and in the validity of these IRT-derived scales.

The strong relationships between the original and the IRT-derived ANX and DEP scales found in this research merit further exploration of other IRT-derived PAI-A scales. However, while the current research suggests that additional exploration of the PAI-A using IRT methods may be fruitful, it is possible that ANX and DEP, as common and related clinical constructs, lend themselves more easily to reduction using IRT; other PAI-A scales may not respond similarly. Nevertheless, one goal in assessment is to produce brief, but high-performing measures, and these results suggest that the IRT-derived ANX and DEP scales, even in the absence of the remainder of the PAI-A, may fulfill this goal. To this end, it may be possible to create an abbreviated

anxiety/depression screener using the IRT-derived version of these two scales, or to use these two scales as individual instruments.

While the results of this study do indicate that the IRT-revised versions perform very well compared to the full-length versions, this study does not provide specific evidence that the IRT-revised versions perform *better* than the full-length scales. The primary benefits of using the IRT-revised versions of ANX and DEP are in reduced client and clinician burden and overall efficiency, rather than an explicit improvement in the overall effectiveness over the full-length scales. Thus, use of the IRT-revisions is preferable for these reasons, but does not directly oppose the use of the full-length PAI-A ANX and DEP scales. Further IRT-based research investigating all PAI-A scales may be instructive as to whether IRT-revisions would be helpful across the PAI-A or if the current results are specific to ANX and DEP.

References

Abela, J. R. Z., & Hankin, B. L. (2008). Depression in children and adolescents: Causes, treatment, and prevention. In J. R. Z. Abela & B. L. Hankin (Eds.), *Handbook of depression in children and adolescents.* (pp. 3-5). New York, NY: Guilford Press.

Achenbach, T. M. (2007). Applications of the Achenbach System of Empirically Based Assessment to children, adolescents, and their parents. In S. R. Smith & L. Handler (Eds.), *The clinical assessment of children and adolescents: A practitioner's handbook.* (pp. 327-344). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213-232. doi: 10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). Reliability, internal consistency, cross-informant agreement, and stability. In T. M. Achenbach (Ed.), *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders (Rev. 3rd ed.)*. Washington, DC: Author.

Anderson, H. E., Jr., & Bashaw, W. L. (1966). Further comments on the internal structure of the MMPI. *Psychological Bulletin, 66*(3), 211-213. doi: 10.1037/h0023623

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation.

Blonigen, D. M., Patrick, C. J., Douglas, K. S., Poythress, N. G., Skeem, J. L., Lilienfeld, S. O., . . . Krueger, R. F. (2010). Multimethod assessment of psychopathy in relation to factors of internalizing and externalizing from the Personality Assessment Inventory: The impact of method variance and suppressor effects. *Psychological Assessment, 22*(1), 96-107. doi: 10.1037/a0017240

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33. doi: 10.1111/j.1745-3992.1997.tb00605.x

Boyle, G. J., & Lennon, T. J. (1994). Examination of the reliability and validity of the Personality Assessment Inventory. *Journal of Psychopathology and Behavioral Assessment, 16*(3), 173-187. doi: 10.1007/bf02229206

Boyle, G. J., Ward, J., & Lennon, T. J. (1994). Personality Assessment Inventory: A confirmatory factor analysis. *Perceptual and Motor Skills, 79*(3, Pt 2), 1441-1442.

Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. [Article]. *Sociological Methods & Research, 21*(2), 230.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011a). IRTPRO 2.1 [Computer software]. Chicago, IL: Scientific Software International, Inc.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011b). *IRTPRO: User's Guide*. Lincolnwood, IL: Author.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2[sup]P[/sup] tables. *British Journal of Mathematical and Statistical Psychology, 59*(1), 173-194. doi: 10.1348/000711005x66419

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. doi: 10.2307/1165285

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*(3), 316-336. doi: 10.1037/0021-843x.100.3.316

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350. doi: 10.1177/014662169301700402

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092-1122. doi: 10.1037/a0021212

10.1037/a0021212.supp (Supplemental)

Copeland, W. E., Shanahan, L., Costello, J., & Angold, A. (2009). Childhood and adolescent psychiatric disorders as predictors of young adult disorders. *Archives of General Psychiatry, 66*(7), 764-772. doi: 10.1001/archgenpsychiatry.2009.85

Costello, E. J., Copeland, W., & Angold, A. (2011). Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when

adolescents become adults? *Journal of Child Psychology and Psychiatry*, *52*(10), 1015-1025. doi: 10.1111/j.1469-7610.2011.02446.x

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*(8), 837-844. doi: 10.1001/archpsyc.60.8.837

Creech, S. K., Evardone, M., Braswell, L., & Hopwood, C. J. (2010). Validity of the Personality Assessment Screener in veterans referred for psychological testing. *Military Psychology*, *22*(4), 465-473. doi: 10.1080/08995605.2010.513265

Dana, R. H., & Cantrell, J. D. (1988). An update on the Millon Clinical Multiaxial Inventory (MCMI). *Journal of Clinical Psychology, 44*(5), 760-763. doi: 10.1002/1097-4679(198809)44:5<760::aid-jclp2270440516>3.0.co;2-c

Davis, K. M., & Archer, R. P. (2010). A critical review of objective personality inventories with sex offenders. *Journal of Clinical Psychology*, *66*(12), 1254-1280. doi: 10.1002/jclp.20722

De Los Reyes, A., & Kazdin, A. E. (2005). Informant Discrepancies in the Assessment of Childhood Psychopathology: A Critical Review, Theoretical Framework, and Recommendations for Further Study. *Psychological Bulletin*, *131*(4), 483-509. doi: 10.1037/0033-2909.131.4.483

Deisinger, J. A. (1995). Exploring the factor structure of the Personality Assessment Inventory. *Assessment*, *2*(2), 173-179.

Devine, D., Kempton, T., & Forehand, R. (1994). Adolescent depressed mood and young adult functioning: A longitudinal study. *Journal of Abnormal Child Psychology:*

An official publication of the International Society for Research in Child and Adolescent Psychopathology, 22(5), 629-640. doi: 10.1007/bf02168942

Dierker, L. C., Albano, A. M., Clarke, G. N., Heimberg, R. G., Kendall, P. C., Merikangas, K. R., . . . Kupfer, D. J. (2001). Screening for anxiety and depression in early adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(8), 929-936. doi: 10.1097/00004583-200108000-00015

Dobson, K. S., & Cheung, E. (1990). Relationship between anxiety and depression: Conceptual and methodological issues. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders.* (pp. 611-632). Washington, DC: American Psychiatric Association.

Dozois, D. J. A., & Covin, R. (2004). The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment.* (pp. 50-69). Hoboken, NJ: John Wiley & Sons Inc.

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7*(4), 435-453. doi: 10.1093/clipsy/7.4.435

Esbjørn, B. H., Hoeyer, M., Dyrborg, J., Leth, I., & Kendall, P. C. (2010). Prevalence and co-morbidity among anxiety disorders in a national cohort of psychiatrically referred children and adolescents. *Journal of Anxiety Disorders, 24*(8), 866-872. doi: 10.1016/j.janxdis.2010.06.009

Fichter, M. M., Kohlboeck, G., Quadflieg, N., Wyschkon, A., & Esser, G. (2009). From childhood to adult age: 18-year longitudinal results and prediction of the course of mental disorders in the community. *Social Psychiatry and Psychiatric Epidemiology, 44*(9), 792-803. doi: 10.1007/s00127-009-0501-y

Frazier, T. W., Naugle, R. I., & Haggerty, K. A. (2006). Psychometric adequacy and comparability of the short and full forms of the Personality Assessment Inventory. *Psychological Assessment, 18*(3), 324-333. doi: 10.1037/1040-3590.18.3.324

Gouge, A. P. (2009). Item response theory analyses of the Personality Assessment Inventory in samples of methadone maintenance patients and university students. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 70*.

Greenbaum, P. E., Dedrick, R. F., & Lipien, L. (2004). The Child Behavior Checklist/4-18 (CBCL/4-18). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment.* (pp. 179-191). Hoboken, NJ: John Wiley & Sons Inc.

Grills, A. E., & Ollendick, T. H. (2002). Issues in parent-child agreement: The case of structured diagnostic interviews. *Clinical Child and Family Psychology Review, 5*(1), 57-83. doi: 10.1023/a:1014573708569

Helmes, E. (1993). A modern instrument for evaluating psychopathology--The Personality Assessment Inventory. *Journal of Personality Assessment, 61*(2), 414-417.

Helmes, E., & Reddon, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin, 113*(3), 453-471. doi: 10.1037/0033-2909.113.3.453

Hoekstra, W. R. (2000). Utility of the Personality Assessment Inventory in adolescents: Discrimination ability with violent versus nonviolent offending incarcerated juveniles. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 60.*

Hoelzle, J. B., & Meyer, G. J. (2009). The invariant component structure of the Personality Assessment Inventory (PAI) full scales. *Journal of Personality Assessment, 91*(2), 175-186. doi: 10.1080/00223890802634316

Hopwood, C. J., Ambwani, S., & Morey, L. (2007). Predicting nonmutual therapy termination with the Personality Assessment Inventory. *Psychotherapy Research, 17*(6), 706-712. doi: 10.1080/10503300701320637

Hopwood, C. J., Baker, K. L., & Morey, L. C. (2008). Extratest validity of Selected Personality Assessment Inventory scales and indicators in an inpatient substance abuse setting. *Journal of Personality Assessment, 90*(6), 574-577. doi: 10.1080/00223890802388533

Hopwood, C. J., Blais, M. A., & Baity, M. R. (2010). Introduction. In M. A. Blais, M. R. Baity & C. J. Hopwood (Eds.), *Clinical applications of the Personality Assessment Inventory.* (pp. 1-12). New York, NY: Routledge/Taylor & Francis Group.

Hopwood, C. J., Creech, S. K., Clark, T. S., Meagher, M. W., & Morey, L. C. (2007). The convergence and predictive validity of the Multidimensional Pain Inventory

and the Personality Assessment Inventory among individuals with chronic pain. *Rehabilitation Psychology, 52*(4), 443-450. doi: 10.1037/0090-5550.52.4.443

Horvath, P. (1992). The MMPI-2 considered in the contexts of personality theory, external validity, and clinical utility. *Canadian Psychology/Psychologie canadienne, 33*(1), 79-83. doi: 10.1037/h0084655

Hsu, L. M. (1994). Item overlap correlations: Definitions, interpretations, and implications. *Multivariate Behavioral Research, 29*(2), 127-140. doi: 10.1207/s15327906mbr2902_1

Hsu, L. M. (2005). Using Critiques of the MCMI to Improve MCMI Research and Interpretations. In R. J. Craig (Ed.), *New directions in interpreting the Millon™ Clinical Multiaxial Inventory-III (MCMI-III™).* (pp. 290-320). Hoboken, NJ US: John Wiley & Sons Inc.

Ingram, R. E., & Siegle, G. J. (2009). Methodological issues in the study of depression. In I. H. Gotlib & C. L. Hammen (Eds.), *Handbook of depression (2nd ed.).* (pp. 69-92). New York, NY: Guilford Press.

Kashani, J. H., Sherman, D. D., Parker, D. R., & Reid, J. C. (1990). Utility of the Beck Depression Inventory with clinic-referred adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 29*(2), 278-282. doi: 10.1097/00004583-199003000-00018

Kearney, C. A., & Bensaheb, A. (2007). Assessing anxiety disorders in children and adolescents. In S. R. Smith & L. Handler (Eds.), *The clinical assessment of children and adolescents: A practitioner's handbook.* (pp. 467-483). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Keenan, K., Feng, X., Hipwell, A., & Klostermann, S. (2009). Depression begets depression: Comparing the predictive utility of depression and anxiety symptoms to later depression. *Journal of Child Psychology and Psychiatry, 50*(9), 1167-1175. doi: 10.1111/j.1469-7610.2009.02080.x

Klein, D. N., Torpey, D. C., & Bufferd, S. J. (2008). Depressive disorders. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *Child and adolescent psychopathology.* (pp. 477-509). Hoboken, NJ: John Wiley & Sons Inc.

Koksal, F., & Power, K. G. (1990). Four Systems Anxiety Questionnaire (FSAQ): A self-report measure of somatic, cognitive, behavioral, and feeling components. *Journal of Personality Assessment, 54*(3-4), 534-545. doi: 10.1207/s15327752jpa5403&4_10

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A New Approach to Integrating Data From Multiple Informants in Psychiatric Assessment and Research: Mixing and Matching Contexts and Perspectives. *The American Journal of Psychiatry, 160*(9), 1566-1577. doi: 10.1176/appi.ajp.160.9.1566

Krefetz, D. G., Steer, R. A., Gulab, N. A., & Beck, A. T. (2002). Convergent validity of the Beck Depression Inventory-II with the Reynolds Adolescent Depression Scale in psychiatric inpatients. *Journal of Personality Assessment, 78*(3), 451-460. doi: 10.1207/s15327752jpa7803_05

Leary, T. (1957). *Interpersonal diagnosis of personality; a functional theory and methodology for personality evaluation*. Oxford England: Ronald Press.

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: Expert ratings and confirmatory factor analyses of the CBCL. *Journal of Consulting and Clinical Psychology, 69*(4), 683-698. doi: 10.1037/0022-006x.69.4.683

March, J. S., & Parker, J. D. A. (2004). The Multidimensional Anxiety Scale for Children (MASC). In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Volume 2: Instruments for children and adolescents (3rd ed).* (pp. 39-62). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

March, J. S., Parker, J. D. A., Sullivan, K., & Stallings, P. (1997). The Multidimensional Anxiety Scale for Children (MASC): Factor structure, reliability, and validity. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(4), 554-565. doi: 10.1097/00004583-199704000-00019

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732. doi: 10.1007/s11336-005-1295-9

Merydith, E. K., & Phelps, L. (2009). Convergent validity of the MMPI-A and MACI scales of depression. *Psychological Reports, 105*(2), 605-609. doi: 10.2466/pr0.105.2.605-609

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066x.50.9.741

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research, 105*(3), 255-264. doi: 10.1016/s0165-1781(01)00317-1

Millon, T. (1994). *Manual for the Millon Clinical Multiaxial Inventory (MCMI-III).* Minneapolis, MN: National Computer Systems.

Mogge, N. L., Steinberg, J. S., Fremouw, W., & Messer, J. (2008). The Assessment of Depression Inventory (ADI): An appraisal of validity in an outpatient sample. *Depression and Anxiety, 25*(1), 64-68. doi: 10.1002/da.20247

Moran, P. W., & Lambert, M. J. (1983). A review of current assessment tools for monitoring changes in depression. In M. J. Lambert, E. R. Christensen & S. S. Dejulio (Eds.), *The Assessment of Psychotherapy Outcomes*. New York, NY: Wiley.

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Morey, L. C. (1996). *An interpretive guide to the Personality Assessment Inventory (PAI).* Lutz, FL: Psychological Assessment Resources, Inc.

Morey, L. C. (1999). Personality Assessment Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment (2nd ed.).* (pp. 1083-1121). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Morey, L. C. (2000). The challenge of construct validity in the assessment of psychopathology. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy.* (pp. 141-171). New York, NY: Kluwer Academic/Plenum Publishers.

Morey, L. C. (2003). *Essentials of PAI assessment*. Hoboken, NJ US: John Wiley & Sons Inc.

Morey, L. C. (2007). *The Personality Assessment Inventory – Adolescent professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Morey, L. C., & Hopwood, C. J. (2006). The Personality Assessment Inventory and the Measurement of Normal and Abnormal Personality Constructs. In S. Strack (Ed.), *Differentiating normal and abnormal personality (2nd ed.).* (pp. 451-471). New York, NY US: Springer Publishing Co.

Morey, L. C., & Hopwood, C. J. (2008). The Personality Assessment Inventory. In R. P. Archer & S. R. Smith (Eds.), *Personality assessment.* (pp. 167-211). New York, NY US: Routledge/Taylor & Francis Group.

Morey, L. C., Warner, M. B., & Hopwood, C. J. (2007). The Personality Assessment Inventory: Issues in legal and forensic settings. In A. M. Goldstein (Ed.), *Forensic psychology: Emerging topics and expanding roles.* (pp. 97-126). Hoboken, NJ: John Wiley & Sons Inc.

Nezu, A. M., Nezu, C. M., Friedman, J., & Lee, M. (2009). Assessment of depression. In I. H. Gotlib & C. L. Hammen (Eds.), *Handbook of depression (2nd ed.).* (pp. 44-68). New York, NY: Guilford Press.

Pine, D. S., Cohen, P., Gurley, D., Brook, J., & Ma, Y. (1998). The risk for early-adulthood anxiety and depressive disorders in adolescents with anxiety and depressive disorders. *Archives of General Psychiatry, 55*(1), 56-64. doi: 10.1001/archpsyc.55.1.56

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

Romain, P. M. (2001). Use of the personality assessment inventory with an ethnically diverse sample of psychiatric outpatients. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 61*.

Rudolph, K. D. (2009). Adolescent depression. In I. H. Gotlib & C. L. Hammen (Eds.), *Handbook of depression (2nd ed.).* (pp. 444-466). New York, NY: Guilford Press.

Ruiz, M. A., & Edens, J. R. (2008). Recovery and replication of internalizing and externalizing dimensions within the Personality Assessment Inventory. *Journal of Personality Assessment, 90*(6), 585-592. doi: 10.1080/00223890802388574

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2).

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*(1), 17-35. doi: 10.2333/bhmk.23.17

Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models.* (pp. 77-107). New York, NY: Routledge/Taylor & Francis Group.

Schlosser, B. (1992). Computer Assisted Practice. *The Independent Practitioner, 12*, 12-15.

Semrud-Clikeman, M., Fine, J. G., & Butcher, B. (2007). The assessment of depression in children and adolescents. In S. R. Smith & L. Handler (Eds.), *The clinical*

assessment of children and adolescents: A practitioner's handbook. (pp. 485-503). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Shaffer, D., Fisher, P., & Lucas, C. (2004). The Diagnostic Interview Schedule for Children (DISC). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment.* (pp. 256-270). Hoboken, NJ: John Wiley & Sons Inc.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(1), 28-38. doi: 10.1097/00004583-200001000-00014

Siefert, C. J., Sinclair, S. J., Kehl-Fie, K. A., & Blais, M. A. (2009). An item-level psychometric analysis of the Personality Assessment Inventory: Clinical scales in a psychiatric inpatient unit. *Assessment, 16*(4), 373-383. doi: 10.1177/1073191109333756

Sinclair, S. J., Antonius, D., Shiva, A., Siefert, C. J., Kehl-Fie, K., Lama, S., . . . Blais, M. A. (2010). The psychometric properties of the Personality Assessment Inventory-Short Form (PAI-SF) in inpatient forensic and civil samples. *Journal of Psychopathology and Behavioral Assessment, 32*(3), 406-415. doi: 10.1007/s10862-009-9165-x

Sinclair, S. J., Siefert, C. J., Shorey, H. S., Antonius, D., Shiva, A., Kehl-Fie, K., & Blais, M. A. (2009). A psychometric evaluation of the Personality Assessment

Inventory–Short Form clinical scales in an inpatient psychiatric sample. *Psychiatry Research, 170*(2-3), 262-266. doi: 10.1016/j.psychres.2008.11.001

Storch, E. A., Storch, J. B., Killiany, E. M., & Roberti, J. W. (2005). Self-Reported psychopathology in athletes: A comparison of intercollegiate student-athletes and non-athletes. *Journal of Sport Behavior, 28*(1), 86-98.

Stulz, N., & Crits-Christoph, P. (2010). Distinguishing anxiety and depression in self-report: Purification of the Beck Anxiety Inventory and Beck Depression Inventory-II. *Journal of Clinical Psychology, 66*(9), 927-940.

Tasca, G. A., Wood, J., Demidenko, N., & Bissada, H. (2002). Using the PAI with an eating disordered population: Scale characteristics, factor structure and differences among diagnostic groups. *Journal of Personality Assessment, 79*(2), 337-356. doi: 10.1207/s15327752jpa7902_14

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology*. (pp. 148-177). Thousand Oaks, CA: Sage Publications Ltd.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 67-113). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Weems, C., & Silverman, W. (2008). Anxiety disorders. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *Child and adolescent psychopathology.* (pp. 447-476). Hoboken, NJ: John Wiley & Sons Inc.

Wetzler, S., & Marlowe, D. (1992). What they don't tell you in the test manual: A response to Millon. *Journal of Counseling & Development, 70*(3), 427-428. doi: 10.1002/j.1556-6676.1992.tb01628.x

Wise, E. A., Streiner, D. L., & Walfish, S. (2010). A review and comparison of the reliabilities of the MMPI-2, MCMI-III- and PAI presented in their respective test manuals. *Measurement and Evaluation in Counseling and Development, 42*(4), 246-254. doi: 10.1177/0748175609354594

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Footnotes

[1] Due to copyright restrictions, item text has been omitted for PAI-A items.

[2] Reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Personality Assessment Inventory™ -Adolescent (PAI®-A) by Leslie C. Morey, Ph.D., Copyright 1990, 1991, 1998, 2007 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.