Item Response Theory Analyses of the Personality Assessment Inventory in Samples of Methadone Maintenance Patients and University Students

By

Albert Patrick Gouge

A DISSERTATION SUBMITTED TO LAKEHEAD UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN CLINICAL PSYCHOLOGY

2008

Your file Votre référence
ISBN: 978-0-494-45690-3
Our file Notre référence
ISBN: 978-0-494-45690-3

Lakehead
UNIVERSITY

OFFICE OF GRADUATE STUDIES

NAME OF STUDENT: Albert Gouge

DEGREE AWARDED: Ph.D.

ACADEMIC UNIT: Psychology

TITLE OF DISSERTATION: An Item Response Analysis of the Personality Assessment Inventory in Normal, Psychiatric, and Methadone Maintenance Treatment

This dissertation has been prepared

under my supervision

and the candidate has complied

with the Doctoral regulations.

Signature of Supervisor

April 24, 2008

Date

Acknowledgements

I would like to thank Dr. Brian O'Connor for his kindness, support, and guidance throughout this process. As well, similar thanks are extended to Dr. Dwight Mazmanian and Dr. Christina Van Barneveld. Many thanks as well to Dr. Charles Netley and Dr. Michael Wesner for their assistance as part of the defense committee. Although she did not play an official role in this dissertation, I must thank Dr. Josephine Tan for her support, mentorship, and guidance throughout my doctoral education.

The completion of this dissertation would not have been possible without the generosity of Dr. Arthur Alterman of the University of Pennsylvania who so kindly provided a large archival data set.

Finally and most importantly I share this accomplishment with my wife Tracy whose generosity, sacrifice, and perseverance allowed me to complete this dissertation and my graduate education. Along with our daughters, Lauren and Leeah, she provided the inspiration and motivation necessary for me to continue through this arduous process. I dedicate this dissertation to the memory of my parents, Catherine and Patty.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation employed a variety of item response theory models and methods to examine the psychometric function of the Personality Assessment Inventory (PAI) in a sample of opioid dependent individuals ($N = 323$) receiving methadone maintenance treatment. The analyses employed both nonparametric and parametric models to examine the PAI scales and items for monotonicity, dimensionality, discrimination, difficulty, information, differential item functioning, and differential test functioning. A large sample of post-secondary students ($N = 919$) were employed as a comparison group for the examination of differential item and test functioning. These analyses resulted in a potential revised version of the PAI for use in methadone maintenance populations and other substance abusing populations. Most scales and subscales were reduced in length by approximately 50% yet retained the majority of the information offered. Many scales and subscales were demonstrated to be multidimensional. Two causal factors were hypothesized with respect to this demonstration of multidimensionality. First, the widespread use of negatively scored items in the PAI likely results in the inclusion of items which are not on the same continuum as the positively worded likely due to artifactual method effects. The second major issue with respect to multidimensionality is the inclusion of symptoms of Axis I and Axis II disorders on the same scale. As well, many other items were shown to offer little in the way of discriminatory power or information. Many scales displayed items with statistically significant differential item functioning (DIF), however, only a few scales were found to have significant differential test functioning. It is suggested that this revised version of the PAI offers an improved alternative scoring method for the assessment of substance abusing populations.

An Item Response Theory Analysis of the Personality Assessment Inventory in a Sample of Methadone Maintenance Patients and a Sample of Post-Secondary Students

The Personality Assessment Inventory (PAI) is a widely used self-report measure of psychopathology and personality. It has been validated for use in a variety of clinical populations employing Classical Test Theory (CTT) methods. There has been little if any examination of the functioning of the individual items and scales of the PAI. Item Response Theory (IRT) and its associated methods of analysis offer an alternative method of psychometric analysis to allow just such an examination. The purpose of this dissertation is to examine the psychometric function of the PAI from an IRT perspective within a diverse sample of methadone maintenance treatment (MMT) patients. Through the identification of the most effective items of each PAI scale or subscale an alternative version of the PAI and or alternative scoring algorithm will be achieved.

The Personality Assessment Inventory (PAI)

The PAI (Morey, 1991) is comprised of 344 questions that make up 4 scales assessing response bias, 11 clinical scales, 4 treatment related scales, and 2 scales assessing interpersonal style. The clinical scales of the PAI include Somatic Concerns, Anxiety, Anxiety-Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, and Drug Problems. The PAI was developed and intended for use in adult populations 18 years of age and over (Morey, 1991). The author designed this inventory for the purposes of providing relevant information to aid in clinical diagnosis, treatment planning, and screening of psychopathology, but not as a measure of normal personality. The PAI manual reports the results of a reading level analysis which indicated that a respondent must have a

minimum fourth grade reading ability in order to complete the inventory. The PAI can be administered through the use of several different paper and pencil forms or computerized software. The authors have recently made a CD-ROM available in which the items are read aloud in a standardized manner for individuals with reading difficulties.

This inventory was constructed using a sequential four-part construct validation strategy which placed emphasis upon both empirical and rational methods (Morey, 1991). The author chose clinical syndromes for assessment based upon both historical importance and contemporary relevance. Raw scores of the PAI scales are converted to T Scores based upon a national U.S. census-matched sample (N = 1000), a large clinical sample (N=1246), and a sample of college students (N = 1051). During its development, the psychometric function of the PAI was examined through a variety of CTT methods to examine such qualities as internal consistency reliability and construct validity. Such methods rely upon statistics that average across individuals and provide little information regarding the contribution or effectiveness of individual items. The author did indicate in the user manual that the PAI was constructed based upon IRT principles; however, no actual IRT analyses were reported.

PAI Inventories of Questionable Acceptability

The PAI contains four validity scales that are intended to aid the clinician in assessing the validity of an administration. These fours scales include the Inconsistency scale (ICN), the Infrequency scale (INF), the Negative Impression scale (NIM), and the Positive Impression scale (PIM). As well, the manual indicates that a valid administration of the PAI can include no more than 17 unanswered items. The ICN scale is intended to assess the consistency of a respondent's answers to items with similar content (Morey,

1991). The scale consists of ten item pairs that query similar item content. The INF scale is intended to identify individual respondents who respond to items in an atypical manner based upon extremely low rates of endorsement in both normal and clinical populations (Morey, 1991). The author hypothesizes that elevations on this scale indicate individuals that have not attended sufficiently to the content of the items of this scale. The NIM scale was designed to identify individuals who respond in manner that gives an exaggerated negative view of the individual. The author suggests that high scores on this scale are possibly indicative of negative self-presentation, malingering, or careless responding. The PIM scale asks the respondent to endorse items that would give a highly favourable impression or asks them to deny minor faults. Again, these items were chosen based upon low endorsement rates in the various samples used for the development of the inventory (Morey, 1991).

PAI Clinical Scales

Somatic Complaints (SOM).

This scale is comprised of three subscales that intend to assess three underlying components of somatoform disorders; Conversion (SOM-C), Somatization (SOM-S), and Health Concerns (SOM-H). Somatoform disorders typically involve symptoms which are typical of organic disorders, but for which no physiological etiology can be determined. According to Morey (1991) the SOM-C scale assesses symptoms associated with conversion disorders such as temporary or episodic paralysis or blindness. The SOM-S subscale was designed to assess the frequency of everyday physical complaints including head aches and back pain. Finally the purpose of the SOM-H subscale is to assess an

individual's preoccupation with health problems, the complexity of such problems, and the extent of their behaviours intended to reduce these problems.

Anxiety (ANX).

The ANX scale and its constituent subscales were designed to examine the affective, cognitive, and physiological symptoms of anxiety, but not the behaviours associated with specific anxiety disorders. That task is assigned to a separate scale and set of subscales, the Anxiety Related Disorders scale. The ANX-A subscale attempts to measure feelings of tension, panic, and nervousness, while the ANX-C subscale is intended to assess worry and rumination. The items of the ANX-P subscale examine the physiological symptoms of anxiety including heart palpitations, sweaty palms, and dizziness (Morey, 1991).

Anxiety Related Disorders (ARD).

The Anxiety Related Disorders scale is comprised of the Obsessive Compulsive (ARD-O), the Phobias (ARD-P), and the Traumatic Stress (ARD-T) subscales which are intended to aid in the diagnosis of various anxiety disorders. The ARD-O subscales targets symptoms of obsessive compulsive disorder as well as personality features. The ARD-P assesses common phobias such as those of heights, enclosed places, and social situations (Morey, 1991), while the ARD-T subscale queries experiences consequential to traumatic events without specifying the nature of the event. The author indicates that an elevation on the ARD-T subscale must be followed up with further questioning to determine if the nature of the event is consistent with PTSD.

Depression (DEP).

The Depression scale of the PAI and its component subscales were designed by the test author to query and measure the main components of the construct. The Cognitive (DEP-C) subscale measures cognitive distortions such as thoughts of hopelessness and helplessness. The Affective (DEP-A) subscale queries feelings of dysphoria and apathy, while the Physiological (DEP-P) subscale examines changes in sleep patterns, appetite, and sexual interest (Morey, 1991).

Mania (MAN).

The author of the PAI designed the MAN scale to measure prototypical signs of a manic episode including disruptions in mood, thoughts, and behaviour (Morey, 1991). The Activity Level (MAN-A) subscale queries symptoms and behaviours such as a reduced need for sleep, psychomotor agitation, and pressured speech. The Grandiosity (MAN-G) subscale examines inflated thoughts of self-esteem, over confidence, and belief in the value of one's ideas. The Irritability (MAN-I) subscale contains items that reflect impatience and high expectations of others. The subscales do not assess for psychotic symptoms sometimes associated with a manic episode.

Paranoia (PAR).

The Paranoia scale was designed to assess symptoms of both paranoid personality disorder and those associated with psychotic paranoia. The Persecution (PAR-P) subscale contains items that deal with severe delusional beliefs. The Hypervigilance subscale is intended to subtly examine an attitude of readiness and wariness (Morey, 1991). The Resentment (PAR-R) subscale queries feeling of resentment, envy, and of being unfairly treated.

Schizophrenia (SCZ).

The Schizophrenia scale examines three core features of schizophrenia: negative symptoms (Social Detachment; SCZ-S), positive symptoms (Psychotic; SCZ-S), and thought disorder (SCZ-T). Positive symptoms addressed by the SCZ-P symptoms include hallucinations, delusions, and bizarre thoughts. The SCZ-S scale is intended to measure symptoms and behaviours associated with either schizophrenia or schizoid personality disorder that include poor interpersonal rapport, flat affect, and poverty of communication (Morey, 1996), According to Morey (1996), the SCZ-T subscale is intended to measure a diverse range of symptoms including tangential speech, thought blocking, and attentional problems.

Borderline Features (BOR).

The BOR scale was designed to measure a broad area of symptoms associated with (but not exclusive to) borderline personality disorder. The Affective Instability subscale (BOR-A) is intended to measure sudden mood changes between states of anxiousness, anger, depression, and irritability without the predictable cyclic nature of bipolar disorders (Morey, 1996). The Identity Problems (BOR-I) subscale queries the respondent regarding his or her sense of self, the idealization and devaluation of others, fear of abandonment, and need for others (Morey, 1996). The Negative Relationships scale (BOR-N) surveys the extent of stormy relationships (chaotic and intense) and the associated feelings of disappointment, distrust, betrayal, and exploitation (Morey, 1996). The Self Harm subscale (BOR-S) was developed to assess an individual's tendency to engage in self harming behaviours without concern regarding the consequences of such behaviours (Morey, 1996).

Antisocial Features (ANT).

This is the second full scale intended to assess symptoms associated with personality disorders. The Antisocial Behaviors (ANT-A) subscale investigates criminal behaviours resulting in theft, destruction of property, and physical violence (Morey, 1996). The Egocentricity (ANT-E) subscale queries the respondent regarding the presence of high levels of self centeredness associated with psychopathy and narcissism. The Stimulus Seeking (ANT-S) subscale is intended to measure the extent to which an individual engages in risky and novel behaviours (Morey, 1996).

Alcohol Problems (ALC) and Drug Problems (DRG).

According to Morey (1996), the ALC scale assesses behaviours and consequences associated with alcohol related disorders. Specific items range from positively worded items querying severe symptoms (e.g., cravings) to negatively worded items regarding total abstinence. The DRG scale queries similar issues related to illicit drug use.

PAI Treatment Consideration Scales

This series of scales is not intended to directly inform diagnosis, but rather to aid the clinician in treatment recommendation by addressing key constructs that have a direct bearing upon the form and process necessary for successful interventions.

Aggression (AGG).

The AGG scale was developed to assess three constructs associated with aggression. The Aggressive Attitude (AGG-A) subscale measures characteristics of individuals hypothesized to contribute to aggression such as the presence of a quick temper or the belief that aggression is an effective method of meeting one's objectives (Morey, 1995). The Verbal Aggression (AGG-V) subscale assesses an individual's predilection towards the expression of verbal anger ranging from sarcasm to loud abusive

language (Morey, 1996). The Physical Aggression subscale addresses the past or present tendency to engage in physically aggressive behaviours.

Suicidal Ideation (SUI).

The SUI scale asks the respondent to endorse or deny specific symptoms and behaviours associated with suicidal ideation such as thoughts of death, the contemplation of suicide, and the currency of such thoughts. The scale does not necessarily indicate the probability that an individual will act on such thoughts.

Stress (STR).

The stress scale queries the respondent with respect to the current presence of life stressors such as problems involving his or her occupation, family, financial status, and relationships. Also measured are major changes that typically result in an increase in perceived stress.

Nonsupport (NON).

The NON scale measures the extent to which an individual perceives a lack of social support by examining his or her interactions with acquaintances, friends, and family.

Treatment Rejection (RXR).

This scale was developed to measure attitudes that are indicative of a lack of motivation towards treatment. Such motivational factors include a willingness to participate in evaluations, a willingness to sacrifice, and openness to new ideas (Morey, 1996).

Interpersonal Scales.

Dominance (DOM) and Warmth (WRM).

The DOM scale was intended to assess the degree to which an individual is controlling, submissive or autonomous with respect to his or her interpersonal relationships (Morey, 1991). Similarly, the WRM scale is designed to assess the extent to which a respondent is empathetic, engaging, rejecting, or distrustful in these same relationships.

Item Development

The items of the PAI were developed based upon the philosophy that the items reflect the phenomenology of the various pathological constructs and personality traits and that the items should be written in a manner that captures the experience of the respondent rather than that of the clinician (Morey, 1991). Empirically derived items that do not reflect construct validity were avoided in the development of this inventory; therefore the author placed special emphasis upon content validity across the range of symptom severity (Morey, 1991). The focus of subscales was decided based upon previous empirical evidence. Item generation was completed by a research team that included clinical practitioners, post-doctoral fellows, and graduate students in clinical psychology (Morey, 1991) based upon the criteria described in the research literature, other inventories, the *DSM*, and clinical experience as guides. However, no other personality tests were examined as exemplars. In the end, 2200 items were developed with at least three times as many items as needed for each scale.

Item Selection

The algorithm used to reduce the 2200 item pool to the final 344 items consisted of a two stage process that involved both conceptual and empirical methods that resulted in two interim versions of the PAI, named Alpha and Beta (Morey, 1991).

The initial stage involved the conceptual evaluation of the item pool accomplished through the completion of three studies assessing the suitability of item content given the intended construct to be measured. In the first study the individual members of the research team rated the item content. In the second study a panel comprised of both professional and lay persons rated the item content. The last study asked external experts from specific areas of mental health to sort the items based upon content. This process reduced the item pool to 776 and constituted the Alpha version of the PAI.

The second stage of this process to choose the most effective items for the final version of the PAI involved two studies. The first study administered the Alpha version of the PAI to a normal sample in order to investigate item distributions, social desirability, gender effects, and the effect of varying item set on malingering and faking. This effort resulted in the Beta version which maintained 597 items. This second to last version of the PAI was then administered to a mixed sample of individuals from both normal and clinical populations in order to examine the psychometric properties of the Beta version. Properties examined included means, standard deviation, item bias (gender, race, or age), item scale correlations and alpha coefficients. Based upon these characteristics the final 344 items were selected.

Psychometric Evaluation of the PAI

The PAI manual states that coefficient alpha can be considered as an estimate of the mean of all possible split half reliabilities and is therefore an appropriate statistic to employ, since the scales of the PAI were intended to be unidimensional (Morey, 1991). Reliability studies reported in the PAI Manual (Morey, 1991) indicated median

coefficient alpha values of .81 for the normal sample, .86 for the clinical sample (Table 1), and .82 for the student sample. The convergent and discriminant validity of the PAI was initially examined through the calculation of correlations between the PAI and other specific and broad psychometric measures of personality and psychopathology (Morey, 1991). These instruments of comparison included the MMPI-2, the NEO-PI, the Beck Anxiety Inventory, the Beck Depression Inventory, the Beck Hopelessness Scale, Mississippi PTSD Scale, Fear Survey Schedule, the Maudsley Obsessive Compulsive Inventory, and the State-Trait Anxiety Inventory. The validation studies were conducting employing community samples, clinical samples, and student samples. In a multitude of analyses each scale and subscale of the PAI was compared to similar and non-similar scales from these other concurrently administered tools to determine the convergent and discriminant validity.

Since its initial development, many of the individual scales of the PAI have been examined for validity in specific populations. For example the validity of the DRG scale was recently examined in a sample consisting of 100 substance using and abusing men and women and 100 non substance-abusing individuals (Kellog et al., 2002). When employed in a sample (N = 200) of normal volunteers, methadone patients, and identified substance abusers the PAI DRG scale correlated significantly with the Addiction Severity Index (ASI) Drug composite score ($r = .81$, $p < .0005$). A significant correlation was also found in the same sample with the ASI Drug Severity rating ($r = .85$, $p < .0005$). The researchers concluded that the DRG scale is a sensitive measure of the negative consequences of drug use, and that the cutoff scores suggested by the test's author (T Score over 70 is consistent with drug abuse and a T Score over 80 is consistent with drug

dependence; Morey, 1991). In this study, 78 % of substance abusers received a T Score over 70, while 99% of those considered to be drug free scored below a T Score of 70. Ruiz, Dickinson, and Pincus (2002) examined the concurrent validity of the PAI ALC scale in a sample of post-secondary student drinkers. The relationships between the ALC scale and patterns of consumption, maladaptive coping, stress, and the results from structured clinical diagnostic interviews (Structured Clinical Interview for the DSM-IV AXIS I Disorders) were examined. It was found that the ALC scale was significantly related to these measures leading these authors to conclude that the ALC scale is a valid indicator of alcohol problems. Parker, Daleiden, and Simpson (1999) studied the relationship between both the PAI DRG and ALC scales and the Addiction Severity Index (ASI) in a sample of chemically dependent patients in a residential treatment centre. These two PAI scales demonstrated good convergent validity with the ASI Alcohol ($r = .49, p < .01$) and Drug ($r = .39, p < .01$) composite scores. Both scales also demonstrated adequate discriminant validity as correlations with other PAI and ASI scales not directly measuring substance or alcohol use were found to be considerably weaker. These results also indicated adequate internal consistency for the ALC ($\alpha = .92$) and the DRG ($\alpha = .78$). The authors indicated that in general the function of the ALC scale was superior to the DRG scale, which they suggest was likely due to the elevated scores with limited range of the DRG scale.

The utility of the PAI in assessing personality disorders has been examined in several different studies. Stein, Pinsker-Aspen, and Hilsenroth (2007) evaluated the criterion and concurrent validity of the PAI in as sample of individuals diagnosed with borderline personality disorder (BPD). The results of this effort indicated that PAI BOR

scale scores were significantly higher amongst those diagnosed with BPD as compared to those who did not carry such a diagnosis and that a T Score of over 70 on the PAI BOR scale correctly classified 73% of the patients diagnosed with BPD. Jacobo, Blais, Baity, and Harley (2007) studied the utility of the PAI in a sample of patients receiving Dialectical Behaviour Therapy, a treatment developed BPD. It was found that scores on the PAI BOR scale were significantly related to SCID-II BPD diagnoses (point biserial correlation = .58, $p < .01$), but not other types of personality disorder diagnoses. With respect to the BOR subscales, it was found that those measuring identity disturbance, self-harm, and negative relationships were uniquely related to the DSM-IV (APA, 1994) criteria for BPD. The BOR total score was also related to the total number of BPD symptoms endorsed on the SCID-II ($r = .63, p < .001$). The authors concluded that the PAI is an effective instrument in the diagnosis of BPD and that a T score above 65 represents an effective cutoff score. Edens, Hart, Johnson, Johnson, and Oliver (2000) investigated the validity of the PAI ANT scale in the assessment of psychopathy in two separate offender populations ($N = 46$ and $N = 55$). Although these authors found high to medium correlations with the Hare Psychopathy Checklist: Screening Version (Hart, Cox, & Hare, 1995; $r = .54, p < .001$) and the Hare Psychopathy Checklist Revised (Hare, 1991; $r = .40, p < .01$), it was found that the PAI ANT scale was related primarily to behavioural symptoms of psychopathy as opposed to the interpersonal and affective aspects. The authors conclude that the PAI ANT scale is an effective dimensional measure of psychopathy as opposed to a categorical measure. Guy, Poythress, Douglas, Skeem, and Edems ((2008) examined the level of agreement between the PAI ANT and SCID-II ASPD scales in a large sample of incarcerated offenders ($N = 678$) and patients

receiving substance-related residential treatment ($N = 667$). Although the two measures were strongly related at the dimensional level, limited levels of categorical agreement were found ($k = .32$) leading the authors to conclude that a cutoff T score of 70 should be considered when employing the PAI as a screening measure for ASPD. A similar study by Wang, Rogers, Giles, Diamond, Herrington-Wang, and Taylor, (1997) compared PAI scores to measures of malingering, suicide risk, and aggression in male inmates. The results indicated moderate correlations between PAI scales (NIM, SUI, & AGG) and criterion variables. The authors indicate that these correlations were likely attenuated due to the restricted ranges (elevated) of PAI scale scores and recommend continued study of the PAI in forensic populations. Bradley, Hilsenroth, Guarnaccia, and Westen (2007) studied the relationship between PAI scores and clinician judgment in the assessment of borderline, antisocial, and obsessive-compulsive personality disorders ($N = 54$). Moderate correlations were demonstrated between the PAI BOR scale and ANT scale and the clinician judgments as quantified by the SWAP BPD and APD scales. Unfortunately, the authors did not report if a similar relationship was found between the PAI ARD-O subscale and the SWAP OCPD scale.

Several studies have examined the psychometric function of the entire PAI in various populations. Schinka (1995) examined the scale characteristics and factor structure of the PAI in a sample of alcohol dependent patients ($N = 301$). This study revealed alpha coefficients (Table 1) that were very similar to those from the standardization sample with the exception of the Alcohol Problems and Treatment Rejections scales. The author suggests that these lower reliabilities are likely due to extremely high means and limited ranges of scores on these scales. Interitem correlations

were also generally found to mirror those of the standardization sample. A factor analysis revealed a close fit with that found in the clinical standardization sample with the exception of a unique factor absent in Morey's analysis. Schinka indicated that this factor represented a dysfunctional interpersonal style associated with alcohol dependence. The author suggests that although these results support the use of the PAI in alcohol dependent populations, further research should examine the psychometric function of the PAI in other specific treatment populations. Karlin, Creech, Grimes, Clark, Meagher, and Morey (2005) examined the psychometric function and utility of the PAI in as sample of chronic pain patients. Internal consistency reliability coefficients were generally found to be acceptable and similar to those found by the test's author with the exception of the NIM, ALC, DRG, and Antisocial Features scales (Table 1). A factor analysis revealed four factors (similar to Morey's results); however, an acting-out behaviour factor identified within the chronic pain group did not include a substance abuse component. The authors hypothesized that this finding may indicate that substance use in chronic pain patients is a maladaptive coping mechanism rather than disinhibited behaviour. The authors conclude that these results support the employment of the PAI within chronic pain populations while highlighting the differences between substance abuse in this population versus general populations.

The reliability, discriminant validity, and construct validity of the PAI was examined in an Australian sample consisting of 131 normals, 30 alcoholics, and 30 individuals diagnosed with Schizophrenia (Boyle & Lennon, 1994). A MANOVA demonstrated significant multivariate main effects for group membership [F (2,378) = 17.307, $p < .001$]. The authors indicated that the normal group received significantly

lower scores overall than the schizophrenia group, who obtained significantly lower scores overall than the Alcoholic group. Reliability coefficients were computed using the Kuder-Richardson Formula 20 (KR_{20}) and found to be generally acceptable for the clinical scales and subscales with most values above .80 (ranging from .58 to .90). However, the following scales demonstrated unacceptable reliability; Infrequency (.60), Positive Impression (.58), Negative Impression (.77), Stress (.79), Treatment Rejection (.76), Dominance (.63), and Warmth (.76). The Alcoholic group and the Schizophrenia group demonstrated significant differences on most scales (at least $p < .01$) with the exception of the Positive Impression, Somatic Complaints, Anxiety, Mania, Schizophrenia, Dominance, and Warmth scales. Further, it was found that the Schizophrenia scale did not significantly discriminate between the two clinical groups. The Alcoholic group also scored significantly higher on the Paranoia scale than did the Schizophrenia group. No schizophrenic patients received a T Score over 90 on the Schizophrenia scale, but 3 Alcoholic patients did receive such scores. The authors indicate that such patterns put into question the construct validity of such scales. The authors conclude that although Somatic Complaints, Anxiety, Mania, Schizophrenia, Dominance, and Warmth scales differentiate effectively between normal and clinical group, they are not effective in discriminating between clinical groups.

McDevitt-Murphy, Weathers, Adkins, and Daniels (2005) examined the use of the PAI in the assessment of PTSD in women ($N = 55$). It was found that the ARD-T subscale ($r = .59$, $p = .01$) and the DEP-P subscale ($r = .64$, $p = .01$) best differentiated those diagnosed with PTSD from those who were not diagnosed. The authors concluded that the PAI is an effective tool in the assessment of PTSD. McDevitt-Murphy, Weathers,

Flood, Eakin, and Benson (2007) further validated the ARD-T subscale in a sample of trauma-exposed college students finding strong evidence of discriminant validity.

Tasca, Wood, Demidenko, and Bissada (2002) examined the use of the PAI in eating disordered populations. To accomplish this task, they examined results from 282 individuals diagnosed with an eating disorder from a total of 478 consecutive referrals to an eating disorder clinic. These authors found that the PAI had acceptable reliability for the clinical and subscales and that the factor structure was quite similar to that reported by Morey. However, Tasca et al. do caution that their finding of an interpersonal factor not evident in Morey's clinical sample could indicate the need for different interpretations for different clinical populations. The authors caution that users of the PAI must pay attention to the population with which they are working. For instance they suspect that treatment-seeking populations may score higher on the NIM scale that would a similar population that was not seeking treatment. They further suggest the utility of norms for specific clinical populations.

The reviewed studies generally support the utility and psychometric acceptability of the PAI in a variety of clinical populations from a CTT perspective. The effectiveness of several specific scales (e.g., ARD-T) has been explicitly examined and found to be adequate. Many of these studies were limited due to elevated scores and limited ranges of scores on the scales of interest when administered to homogeneous clinical populations. As well, specific weaknesses in psychometric function and differences in factor structures of the PAI tended to be demonstrated in alcohol dependent and substance dependent populations. Interestingly, given the decision to sometimes include symptoms of both Axis I and Axis II disorders on the same scale or subscale, most studies reported only

moderate correlations between PAI measures of personality disorders and other measures of such symptoms and behaviours. However, despite specifically searching for previous IRT studies of the PAI none were located. An examination of the effectiveness of the PAI at the item level and the possibility of differential item functioning in different populations would seem an appropriate goal based upon this review.

Development of a New Version of the PAI for Use in Specific Clinical Populations

The concept of packaging the PAI for use in a specific population is not an original concept. Such an effort has been performed in the area of forensic assessments. The publishers of the PAI offer a unique version of the PAI (including specific norms) for use in this population called the PAI-CS. According to the publishers the goal of this special version of the PAI is threefold:

1. The identification of risk of misconduct while incarcerated.
2. Assess the psychosocial needs of the individual.
3. Estimate the individual's response to incarceration and rehabilitative programming.

This version of the PAI contains experimental scales that were developed with the corrections normative data. These scales include an Addictive Characteristics scale intended to identify individuals at risk for substance abuse and a validity index specific to corrections environments.

Substance Dependence, Opioids, and Methadone Maintenance Treatment

The DSM-IV-TR (APA, 2000) divides substance use disorders into substance abuse and substance dependence. The twelve month prevalence rates in the U.S. for substance dependence have been estimated at 1.1% of the population (Kessler et al.,

1994). Substance dependence is associated with increased severity of symptoms and consequences as compared to substance abuse.

Substance abuse refers to significant substance use that results in problems for the individual but does not result in addiction or withdrawal symptoms. DSM-IV-TR (APA., 2000) criteria for Substance Abuse requires that the individual has engaged in a maladaptive pattern of substance abuse that has resulted in clinically significant impairment or distress as indicated by at least one of four main symptoms in a 12-month period.

1. Due to substance use, the individual is unable to meet his or her role obligations (e.g., work, school, or home).
2. Repeated substance abuse in situations that are dangerous such as driving automobile or operating machinery.
3. Repeated legal problems related to substance use.
4. Continued substance use despite negative interpersonal consequences due to or exacerbated by substance use.

In order to receive a diagnosis of Substance Abuse the individual must have never met criteria for Substance Dependence.

According to the DSM-IV-TR (APA., 2000) Substance Dependence refers to a maladaptive pattern of substance use that results in significant impairment or distress that includes at least three major symptoms within a twelve month period. These symptoms include:

1. Tolerance as evidenced by the need to increase amount of the substance used to achieve the desired effects or reduced effects with the continued use of the same amount of the substance.
2. Withdrawal as characterized by the withdrawal syndrome for the specific substance or the use of the substance to avoid the symptoms of withdrawal.
3. The individual uses more of the substance over a longer period of time than was intended.
4. The individual consistently desires to or unsuccessfully attempts to reduce or eliminate use of the substance.
5. The individual spends an inordinate amount of time in efforts to acquire, use, or recover from use of the substance.
6. The individual reduces or gives up engaging in importance social, occupational, or recreational activities because of substance abuse.
7. The individual continues to engage in substance use despite the knowledge that his or her substance abuse is likely to worsen a persistent physiological or psychological problem.

Specifiers for this disorder include with physiological dependence or with psychological dependence.

Concurrently Disordered Populations.

Individuals who suffer from both a major psychiatric disorder and a substance use disorder (SUD) are often referred to as concurrently disordered or dually diagnosed. It is estimated that at least 50% of individuals with a mental disorder suffer from a substance use disorder (Torrey et al., 2002). According to Mueser, Noordsky, Drake, and Fox

(2003), in their treatment manual for dual disorders, the prevalence varies from population to population (20 to 65% of populations with a mental health disorder). They suggest that assessment methods, diagnostic criteria employed, the clinical setting, and the demographics of the population being examined may affect such differences. A concurrent substance use disorder increases negative outcomes such as hospitalizations, incarcerations, homelessness, victimization, and Hepatitis C (Torrey et al., 2002). Not only does the presence of a concurrent disorder increase the difficulty of treatment, but also it increases the difficulty of diagnosis several fold due to overlapping symptoms.

It has been shown that the presence of major psychiatric disorders complicates the treatment of SUDs. Haller, Miles, and Dawson (2002) used the MCMI-II to divide a sample of substance abusing women into three groups based on the severity of their psychiatric symptoms. It was found that women with severe psychiatric symptoms, unstable mood, and interpersonal deficits were less likely to complete treatment. The authors suggest that early diagnosis of psychiatric disorders is an important early step towards effective treatment of SUDs.

Calhoun, Sampson, Bosworth, Feldman, Kirby, Hertsberg, Wampler, Tate-William, Moore, and Beckham (2000) surveyed the use of illicit substances and the validity of self-reported drug use among treatment seeking veterans referred to a clinic specializing in the treatment of Posttraumatic Stress Disorder. The Clinician Administered PTSD scale was used to diagnose PTSD and current drug use was confirmed by urinalysis. It was found that PTSD predicted greater use of marijuana and depressants rather than stimulants. It was also found that self-reported drug use was generally accurate as verified by urinalysis. Ouimette, Brown, and Najavits (1998)

conducted a review of the empirical literature regarding comorbid PTSD and SUD. Based upon this review Ouimette et al. make several recommendations concerning the treatment of substance abusers. They recommended that all substance abusers should be screened for PTSD and referred for specialized concurrent treatment if available. As well, they report that the presence of PTSD is a predictor of more serious substance abuse difficulties and poor response to treatment. Such studies highlight the importance of early accurate diagnosis in these complicated populations and the importance of scales that assess psychopathology in substance abusing populations.

Opioid Dependence.

The DSM-IV-TR (APA., 2000) defines Opioid Dependence as a maladaptive pattern of opioid use that results in significant impairment or distress that includes at least three of a group of symptoms consisting of tolerance, withdrawal, a need for increasing amounts of the drug, a persistent desire or unsuccessful effort to reduce the amount of drugs used, a great deal of time spent using, acquiring or recovering from the use of drugs, interruption of important duties and life roles, and the continued use of the drug despite the knowledge of its deleterious effects on one's life.

Commonly abused opioids include heroin, morphine, and codeine. Commonly abused pain medications include Tylenol III, Percocet, Percodan, Oxycontin, and Oxycodone. Opioids are a highly addictive group of drugs that result in physical and psychological dependence involving intense physical withdrawal. Due to the intense symptoms of withdrawal experienced by opioid addicts they need to maintain their daily supply and often engage in criminal activities (Mueser et al., 2003). In addition to opioid dependence, patients of methadone maintenance treatment centers are often

polysubstance abusing or dependent. Other substances of abuse often include cannabis, cocaine, benzodiazepines, and alcohol (Mueser et al., 2003).

Methadone Maintenance Treatment.

Methadone Maintenance Treatment (MMT) is an agonist therapy for the treatment of opioid dependence. Methadone is a long acting synthetic opioid that is prescribed by a specially licensed physician and is taken daily. Once stabilized the individual no longer experiences a high or a euphoric experience from methadone or from any opioid. As well, due to the long acting nature of methadone the individual should not experience withdrawal as long as they receive his or her daily dose.

Significant empirical evidence exists to support the contention that MMT populations have a high rate of concurrent psychiatric diagnoses. Cacciola, Alterman, Rutherford, McKay, and Mulvaney (2001) studied the relationship of comorbid psychiatric disorders to pre-treatment status and to outcomes in a sample of 278 MMT patients. Cacciola et al. found that 75.2% of this population had a concurrent Axis I or Axis II disorder. It was found that a comorbid psychiatric disorder did not predict poorer outcomes in opioid addiction but did predict poorer psychosocial and medical outcomes. The presence of an Axis II personality disorder did predict higher rates of dropout from treatment. Krause, Degkwitz, Kuhne, and Vertheine (1998) studied 350 opioid addicted individuals in Hamburg Germany to examine the rate of comorbid mental health disorders as defined by the ICD-10. This study found that 55% of the subjects suffered from a concurrent psychiatric disorder not including personality disorders. Kaye, Darke, and Finlay-Jones (1998) reported that the prevalence of Antisocial Personality Disorder

in populations of IV drug users was 35-61 %. These findings indicate that concurrent Axis I and Axis II disorders are commonly present in MMT patients.

The assessment and treatment of MMT patients is further complicated by other addictive behaviours such as gambling. Ledgerwood and Downey (2002) studied a population of MMT patients for the coexistence and consequences of problem gambling. They examined 62 methadone patients using South Oaks Gambling Screen (SOGS, Leiseur & Blume, 1987). They found that 17.7% of the methadone patients studied met criteria a pathological gambling and that the presence of a gambling disorder increased the likelihood of cocaine use and the likelihood of treatment dropout.

According to the Methadone Best Practices manual published by the Ontario Government, untreated opioid addiction results in increased rates of criminal activity, medical services, lost productivity, HIV, and Hepatitis C for both the user and children of users. Risser et al. (2001) examined the mortality rate of opioid abusers in methadone treatment and those not receiving treatment in Vienna, Austria. It was found that those in methadone treatment had a significantly lower mortality rate (12.1 %) as opposed to those in opioid related emergencies (48.8%). Hogan (1998) reviewed the literature regarding the social environment of children of cocaine and opioid abusers and the long-term developmental outcomes for these children. Although in general there is limited research in this area the authors do conclude that infants exposed to heroin parentally suffer from elevated levels of tension and anxiety, while similar exposure to cocaine is associated with physical difficulties such as low birth weight and size. Other studies reviewed by these authors suggest that children of substance abusers are at greater risk for diseases related to neglect (e.g., infectious & nutritional diseases). Evidence suggests

that children of drug dependent parents suffer from deficits in developmental progress, socially adaptive behaviour, social development, and cognitive development. Hogan (1998) reported that the most compelling evidence regarding the negative effects of parental drug abuse is the high likelihood that the children will in turn suffer from substance use disorders as well.

The evidence reviewed suggests that opioid dependent populations have a high prevalence rate of concurrent psychiatric disorders, which in turn predict poorer treatment compliance and poorer psychosocial outcomes. This evidence highlights the need for improved psychosocial outcomes in MMT populations and the importance of improving standard psychological instruments for use in such a specific population.

Item Response Theory

Traditionally, the development and psychometric examination of measures of personality and psychopathology has been based upon Classical Test Theory (CTT). This theory requires the adoption of a conceptual model in which a relationship between constructs is theorized. From such a conceptual model, a measurement model is derived in which the variables to be measured are chosen. In CTT a test score is assumed to be a continuous variable and it is assumed that an individual's observed score is equal to his or her true test score plus error. In CTT this standard error measurement is assumed to be equal across all scores in the population. CTT attempts to measure reliability and validity using several different approaches. A test is considered reliable if similar results are obtained over time (or across forms or raters); while validity refers to the extent to which a scale actually measures the construct for which it is was designed. In CTT, it is assumed that a longer test is more reliable than a shorter test based upon the Spearman

Brown prophecy formula (Embretson, 1996), under the assumption that all items are measuring the same construct. In CTT, the meaning of a score is obtained by comparing the rank of the individual's score in a similar population. This relies upon the transformation of raw scores to standardized scores (e.g., T score) and or percentiles. According to Steinburg and Thissen (1996) traditional methods and statistics used in psychological test construction (e.g., coefficient alpha & item-total correlations) are inadequate to properly examine the complex data measuring psychopathology. For example, these authors indicate that IRT methods provide the researcher with the ability to examine the sources of item covariance and to more accurately estimate internal consistency for a scale with demonstrated construct validity.

Item Response Theory (IRT), despite the complexity of its mathematics, is essentially a theory regarding the process that occurs when an individual responds to a test item (Thissen & Steinburg, 1988). Item Response Theory refers to a group of modern measurement models and was initially developed in the late 1960s. IRT (or latent trait theory) assumes that psychological constructs cannot be directly measured but must be inferred from an individual's responses to items on a scale (Meijer & Baneke, 2004). Unidimensional IRT models are based upon the assumption that the responses to all items on a particular scale depend upon a single continuous underlying variable (Thissen & Steinburg, 1988). An individual's response to an item depends on his or her level of the underlying trait, the effectiveness of the item, and possibly the population to which the individual belongs (in circumstances of DIF). Such models assume that n individuals who possess more of a trait are more likely to endorse an item which measures the trait than individuals who possess less (Santor & Ramsay, 1998). According to Embretson (1996)

IRT varies from CTT in several fundamental ways. For instance, the standard error of the measurement in IRT varies across individual scores and generalizes across populations. This is opposed to the conceptualization of SEM in CTT, which proposes that SEM can be applied to all scores within a population. Other differences include IRT contentions that short tests can be more reliable, that varying levels of test difficulties across persons is preferable when comparing multiple forms of a test, and that unbiased estimates of item properties can be obtained (Embretson, 1996).

Most commonly employed IRT models (nonparametric & parametric) maintain three basic assumptions: monotonicity, unidimensionality, and local independence. Monotonicity is a central facet of most item response models (Junker & Sijtsma, 2000). The concept of monotonicity that underlies these models postulates that if a test item measures a latent trait (θ), then an individual with a higher level of θ is more likely to score higher on that item. Unidimensionality implies that a central factor or facet exists within a scale. A definition of the strong assumption of local independence states that in a subpopulation in which latent traits take fixed values, the responses to items are conditionally independent (McDonald, 1999, pg. 255). From this strong principle, the weak principle of local independence is inferred, which states that pairs of items are uncorrelated in a subpopulation in which the latent traits are fixed (Embretson, 1996).

Nonparametric IRT (NIRT) Models

According to Ramsay (2000) the development of NIRT models was an important progression in modern statistics. Rather than stipulating a specific function defined by a set number of parameters and then estimating these parameters from the data, NIRT models allows for the direct estimation of the function (e.g., kernel smoothing).

Parametric Item Response Theory (PIRT) models are limited by their assumption that the limited number of parameters in the model can accurately describe the data. Ramsay (2000) points out that despite the number of parameters employed, situations still arise in which the model does not deliver the flexibility required to accurately represent the data. On the other hand too many parameters results in the over fitting of more simple data. The accuracy of the estimates of the parameters is dependent upon the correctness of the parametric model chosen. Another problematic assumption of PIRT models is that these parameters are accurate for the entire range of the latent trait, which according to Santor and Coyne (1997) is highly unlikely especially in the severe ranges of interest to clinical researchers.

Meijer and Baneke (2004) suggest several advantages to using an NIRT model in the analysis of personality data. The primary advantage cited by these authors is that NIRT methods do not force the data to fit to a logistic model. This allows for the examination of the structure of the data and is especially useful in the development of a test when choosing items from a pool of possible choices. For example, Nowak, Robertson-Nay, Strong, Bucceria, and Lejeuz (2003) used NIRT (in conjunction with factor analysis) to create a psychometrically sound eating disorder screen for college students, consisting of 7 items from an original pool of 91 items. Using NIRT, these researchers were able to arrive at a small number of items that discriminated over a wide range of symptom severity. NIRT models are especially useful for determining the range of severity of a population in which an item best discriminates. These authors suggest that NIRT models should be used routinely in the initial analysis of data and in determining which parametric model might be useful. According to Santor and Ramsay

(1998), when assessing the measurement of psychopathological constructs, NIRT models are of great utility in determining the effectiveness of options and items at different levels of trait severity, distinguishing true group differences from group differences due to item bias (or differential item functioning), and assessing scale discriminability. Junker and Sijtsma (2001) suggest three primary motivations for the use of NIRT.

1. NIRT can be employed to examine communalities between PIRT and NIRT models through the identification of features (local independence, monotonicity of item response functions, unidimensionality of the latent trait)
2. NIRT models are more flexible and can be used to assess violations of local independence due to nuisance variables, examine differential item functioning, provide context in which to develop methodologies to establish the number of dimensions underlying a test, and to function as an alternative to PIRTs in tests of fits.
3. NIRTs are applicable in situations where there are small numbers of subjects and small numbers of items. NIRT models can also identify groups of items that constitute subscales when a single unidimensional scale does not exist.

Monotonicity is an underlying assumption of IRT that can be explored using an NIRT model. Rather than simply assuming that as an individual possesses "more of the underlying trait" they are more likely to score higher on any particular item, this can actually be assessed using a NIRT. Examination of Option Characteristic Curves (OCCs) allows for the determination of the discriminability of the individual options and over what range of the latent trait the options discriminate amongst individuals. By examining

OCCs it is possible to determine the level of trait severity at which a specific symptom is likely to be observed.

Meijer and Baneke (2004) conclude that NIRT analyses offer information difficult to gain through PIRT, are useful to explore the data (by staying "close" to the data), and can be employed to identify items that may require the use of more complex parametric models such as 3PLM or 4PLM. They further suggest that NIRT should be automatically performed in the initial analysis of a scale, can be useful in the examination of the structure of a scale, and can be useful in the selection of effective items for a specific population.

Santor, Ramsay, and Zuroff (1994), following their NIRT examination of the Beck Depression Inventory II made several key suggestions regarding the employment of NIRT methods. First, they noted that the psychometric properties of scales are generally summarized using item-total correlations or reliability coefficients, when in contrast, IRT models examine how responding to an item varies as a function of dimension. They also state that response characteristic curves function as an estimate of item effectiveness, which may be more informative than such traditional measures. Santor et al. noted that NIRT methods are highly effective in assessing the weights assigned to options. For example, they hoped to examine gender bias but contended such mean differences do not prove group bias. Finally, they point out that parametric models do not specifically allow for significant departures from model assumptions.

Santor and Coyne (1997) point out disapprovingly that attempts to improve psychometric instruments employed for screening purposes invariably attempt to a adjust cut off scores as opposed to determining the effective items and disposing of the

ineffective. Another important feature of NIRT modeling is that fact that it is unaffected by reverse scoring as is often found in the PAI.

Some theorists suggest that NIRT is better suited than PIRT for the development of tests of personality and psychopathology, where the manner in which individuals choose their response is less well understood than in cognitive or educational testing and the choice of an appropriate function difficult. Meijer and Baneke (2004) argue that the benefits of employing NIRT include useful easily interpreted information, easy to use software, and examining the data as it exists rather than forcing it to fit a specific function. NIRT is important in assessing the lack of fit of a selected model and the data. According to Stout (2001), an NIRT examination prior to PIRT analysis should not be viewed as a suggestion but as a requirement in that such an examination allows the researcher to determine the appropriate PIRT with which to follow.

NIRT Estimation of Option and Item Characteristic Curves with Testgraf.

Testgraf is a NIRT software program designed to facilitate the development and evaluation of multiple-choice tests including psychological scales (Ramsay, 2000). The software is available at no cost from the developer. The use of Testgraf does not require specific computer or statistical knowledge. As in most IRT models, Testgraf assumes the scale in question is measuring a single underlying trait. The output from Testgraf is graphical in nature. Typical plots are the option characteristic curve (OCC) and the item characteristic curve (ICC) also known as item response functions (IRF). Testgraf can also be employed to investigate Differential Item Functioning (DIF) between groups. Typically, DIF is used to determine if two different groups (e.g., males or females)

systematically respond differently to items on a scale. This can also be useful to determine if a scale can be used in different cultures.

Testgraf estimates item response functions employing a kernel smoothing technique which does not assume that the function takes a logistic form nor does it force the data to fit such a function. Since NIRT models do not assume that an item will function similarly throughout the continuum of severity of the underlying trait they offer the opportunity to examine how the model and the item functions in different regions of theta.

Gaussian Kernel Smoothing.

According to Ramsay (2000) kernel smoothing is a simple, computationally efficient method (on average 500 to 1000 times faster that a typical software employing a parametric model) of directly estimating a function. He argues that it is especially useful when there is no substantive reason to suppose that the function will display characteristics of a known parametric function.

Testgraf ranks individuals in the sample based upon his or her standardized test score. The values of the OCCs are usually estimated at 51 (up to 101) equally spaced intervals using a nonparametric (Gaussian) kernel smoothing technique that employs local weighted averaging. Local weighted averaging indicates that an average is computed from the values that lie closest to each evaluation point with more weight being given to those points closer in value to the evaluation point.

Option Characteristic Curves.

The main purpose of Testgraf is to graphically illustrate the relationship between the likelihood that an individual will endorse the available options for each item of a

scale and the level of the unobservable underlying trait in the individual (Ramsay, 2000). This relationship is graphically portrayed in an Option Characteristic Curve (OCC) as demonstrated in Figure 1. According to Santor and Ramsay (1998) the OCC from an ideal item will display distinct ranges of severity where each option is most probable in an orderly manner and where changes in probability are rapid. As seen in Figure 1, individuals who score in the bottom two standard deviations are much more likely to choose option 0. As an individual's overall total score rises, they become more likely to choose options with higher values, reaching the point at which those who score highest on the underlying trait are likely to choose the option scored 3.

Item Characteristic Curves in Testgraf.

An item characteristic curve (ICC) graphically demonstrates the relationship between the underlying trait and the item response chosen demonstrating the probability of an individual choosing a particular response given his or her level of the underlying trait. The ICC will graphically demonstrate the degree of monotonicity of the item. As seen in Figure 2, the ICC produced by Testgraf uses vertical bars to indicate the 95% confidence limit for the curve at that level of the specific trait and are referred to as point wise confidence limits (PCL). Given a sufficient sample size, typically at least 100 respondents (Ramsay, 2000), these bars can be used to estimate the effectiveness of the item at different levels of the underlying trait. In this example the bars indicating the PCL are all similar in length indicating that the 95% confidence limits are relatively stable across the continuum of the specific underlying trait. This would provide evidence of the utility of this item across severity levels.

Mokken Scaling Analysis

The Mokken model (Mokken, 1971) is a nonparametric IRT model originally developed to analyze dichotomous data. It consists of two separate models; Monotone Homogeneity (MH) and Double Monotonicity (DM). The DM model parallels the Rasch model also known as the 1 parameter logistic model (PLM), while the MH model more closely parallels the Birnbaum model also known as the 2 Parameter Logistic Model (2PL). Molenaar (1991) extended the Mokken model for the purpose of analyzing polytomous data and this model is useful in examining similar data to the Graded Response Model (Samejima, 1967).

The Mokken model maintains several assumptions regarding the structure of the scale. It assumes the data are monotone, homogeneous, and locally independent. These assumptions are similar to those of most IRT models, both parametric and nonparametric models. Mokken scale analysis is an ordinal unidimensional measurement model based upon Guttmann scale IRT (van Schuur, 2003). The focus of this discussion will be the probabilistic Monotone Homogeneity model. This model was originally devised to examine dichotomous data similar to the 1PL, but was later extended to include polytomous data. According to van Schuur the key benefits of Mokken scaling analysis include distribution parameters, a probabilistic rather than deterministic nature, and a demonstrated ability to perform well with scales that contain only a few items. Mokken scaling was designed to determine when it is appropriate to summarize a set of items by their cumulative score (Alterman, Cacciola, Habing, & Lynch, 2007).

Essentially Mokken scale analysis compares the scale in question to a perfect Guttmann scale. Therefore an adequate understanding of Guttmann scaling is required. The Guttmann scale is cumulative where each subject's item scores are accumulated (van

Schuur, 2003). It is contended that in a Guttmann scale an individual will generally endorse items which indicate less severe symptoms than the most severe item they have endorsed. For example, if an individual has already endorsed being told by others they have a drug problem, they should endorse using drugs. A perfect Guttmann scale predicts both the makeup and frequency of specific response patterns. These patterns conform to the assumptions of the Mokken model, which are very similar to those of most IRT models. The Mokken model assumes monotonicity, homogeneity, and stochastic independence. Although Mokken scale analysis does not expressly examine local independence, its predictions against which the scale in question is compared, are based upon such an assumption. Therefore, if a scale does compare favourably to the model predictions, acceptable levels of local independence are inferred.

For example, in Table 1, response pattern #1 represents a response vector in which an individual does not positively endorse the most severe item (Item #6) and therefore does not endorse any of the less severe items. Response pattern two indicates that such a respondent endorses the least severe response option to the most difficult question and doest not respond positively to the easier items. The frequency column indicates the expected number of individuals (out of 100) who would respond in such a manner assuming local stochastic independence.

A Mokken scale analysis of a polytomous scale would involve the construction of a data matrix for each possible pair of items. Such a matrix would include a column for each data step which would be recoded into dichotomized pairs. From this matrix it is possible to determine the number of Guttmann violations, which are then weighted and summed. The coefficient of homogeneity (also known as Loevinger's *H*) can be

calculated by comparing the observed weighted errors to the weighted predicted errors. A cut off score of .30 has been well established for this scalability coefficient. As well a *Z* statistic is calculated to determine the significance of the finding.

Sijtsma, Debets, and Molenaar (1990) explained the calculation of the scalability coefficient H for pairs of items, items, and the scale as a whole in the case of a polytomous scale. They let *i* and *j* represent two items of a scale, with g representing a response option to item *i* and *h* representing a response option to item *j*. In this situation, option g for item *i* represents a more severe symptom than option h of item i. In this case, the endorsement of option *g* but not *h* represents a Guttmann error. O_{ij} is defined as the sum of observed errors, while E_{ij} represents the total sum of expected errors. Therefore, for two items *i* and *j*:

$$H_{ij} = 1 - O_{ij} / E_{ij} . \qquad (1)$$

If the observed and expected sums of errors are determined for all pairs of item then the item *H* and scale *H* coefficients are calculated by:

$$H_i = 1 - \sum_{j \neq i}^{k} O_{ij} / \sum_{j \neq i}^{k} E_{ij} \text{ , and} \qquad (2)$$

$$H = 1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i=1}^{k} O_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i=1}^{k} E_{ij}} \qquad (3)$$

The Search Strategy in the Mokken Scaling Procedure.

The Mokken scaling procedure is a unique "bottom up" feature that is unlike most other scaling or data reduction procedure (for example factor or reliability analyses). This search procedure is a highly effective test construction tool (Molenaar, 2001). Rather than examining the entire scale and eliminating the most ineffective item, this method searches for the pair of test items with the highest scalability coefficient above 0.30 and a Z statistic above the cut off of 1.65. In the event of ties the most difficult items are chosen. The next best item is selected and the analysis is repeated to determine the acceptability of the overall scale and item coefficients. This process continues until a last added item does not display a coefficient of homogeneity above the 0.30 cut off. If at least two items are eliminated, the remaining items are searched for another pair of items that meet the minimum standards. If such a pair is determined the process continues to establish a second scale. When a scale is constructed employing the Mokken method the respondents can be ordered based upon their test score.

Parametric Item Response Theory (PIRT)

PIRT models rely upon the use of logistic response functions and assume that response curves can be estimated from specific mathematical functions (Santor & Coyne, 1997). According to Uttaro and Lehman (1999) PIRTs offer powerful methods that afford

the user several specific advantages over classical test theory methods such as independent item parameters and latent trait levels, expression of the model at the level of the item rather than total score, the ability to examine differential item functioning, and the equating of scores across forms or sets of items. These authors also point out that the model is expressed at the item level, rather than the test score level which allows for the assessment of the effectiveness and information offered by a specific item. PIRT methods also allow for the assessment of DIF and DTF, the creation of short parallel forms, and the equating of scores from different forms. According to Uttaro and Lehman (1999) total scale information and reliability vary directly indicating that items which offer little information do not contribute meaningfully to the overall level of reliability. The use of PIRT methods in test construction allows for the calculation of item statistics that are independent of the sample from which they were derived and estimates of ability that are independent of items. IRT operates at the level of the item allowing greater flexibility in test development and test scoring.

The Graded Response Model.

Samejima (1969) developed the graded response model (GRM) describing it as a family of mathematical models used to examine polytomous ordered data ranging from letter grades to likert type scales. The GRM is an extension of the 2 parameter logistic (2PL) model for dichotomous items to the polytomous case. The most common form of polytomous items found in the measurement of psychopathology and personality employ Likert-type scales with four or five response options. In this model a two stage process occurs in which the categories are artificially dichotomized and subsequently subtracted in order to determine the probability that an individual will respond in a certain category.

GRM defines the probability that a respondent with a certain trait level will choose a particular option

Item Parameters in the GRM.

In the GRM, Operating Characteristic Curves are calculated that represent the probability of an individual's response to an item falling in or above a given category (Embretson & Reise, 2000). In this model, each item has one less such curve than the number of available options. As in the 2PL, there are two different types of parameters that are computed for each item of a scale. The discrimination parameter (*a*) refers to the ability of the item to discriminate between respondents and is equal to the slope of the Operating Characteristic Curve at its inflection point. In this model, the difficulty parameters (b_1, b_2, $b_{3...}$) refer to the level of the underlying trait at which a respondent has an equal chance of choosing between two options.

In order to calculate response probabilities each item is treated as a series of dichotomies and the 2PL model is estimated for each dichotomy with the same discrimination parameter for each option. The following equation is the logistic form for the calculation of operating characteristic curves (Childs & Chen, 1999; Reise & Yu, 1990).

$$P_{jk}^{*}(\theta) = \frac{1}{1+\exp[-\alpha_j(\theta-\beta_{jk})]} \quad , \qquad (4)$$

Where $P_{ij}^{*}(\Theta)$ is the probability of endorsing option *k* or higher on item *j*, Θ is the latent trait, *j* is the item, α_j is the common slope for all options for item *j*, and β_{jk} is the threshold for option *k* on item *j*,

Following the estimation of the 2PL operating characteristic curves, the category response probabilities are computed as follows (Childs & Chen, 1999; Reise & Yu, 1990).

$$P_{jk}(\theta) = P^{*}_{jk}(\theta) - P^{*}_{j(k+1)}(\theta) \ , \qquad (5)$$

With the following constraints:

$P^{*}_{j0}(\theta) = 1$,

$P^{*}_{j,m+1}(\theta) = 0$,

Where m is the number of response categories minus 1.

And for any value of k,

$P^{*}_{jk}(\theta) - P^{*}_{j,k+1}(\theta) \geq 0$.

This allows the construction of Category Response Curves for each item, where the inflection point of the curves represent the point at which a respondent has a 50% chance of responding above a response category (Embretson & Reise, 2000). The *a* parameter for an item is not strictly viewed or interpreted as a discrimination parameter but is used in conjunction with the difficulty parameters to compute information curves through which the utility of an item is assessed.

Information.

In statistical terms, information is generally defined as the reciprocal of the precision with which a parameter can be estimated (Baker, 2001). Such precision is measured by the variability of the estimates of the parameter (Baker, 2001). With respect

to IRT analyses, the measure of interest is the individual's level of the underlying trait (θ), therefore information can be computed as the reciprocal of the variance of the estimates of the ability levels of the respondents (Baker, 2001). When used with respect to IRT measurement specifically, information refers to the precision or the effectiveness of an item or scale in differentiating between respondents (Reise, Ainsworth, & Havilland, 2005). Item information refers to an item's precision, while total scale information is computed by summing the information provided by each item of the scale and is intended to demonstrate the utility of the scale as a whole. The higher the information at any point, the lower the Standard Error, therefore eliminating items with little information and high levels of error reduces the ratio of error to information. For the GRM, information is related to the discrimination parameter and the range of the of the difficulty parameters. In the GRM, information is usually greatest for items with high discrimination parameters and a narrow range of difficulty parameters. Significant advantage of polytomous IRT models is the information and measurement precision can be estimated at the item option, option, scale, and test level. Although there are no definitive thresholds for identifying low discrimination parameters, it is generally considered that any *a* parameter below 1 is low. A major benefit of IRT and the computing of information curves is the knowledge of where information is highest within the range of theta for a particular item or for the total scale. This knowledge can prove valuable to test developers in choosing items to differentiate respondents within a specific range of severity of the underlying trait. For instance, if developing a test for use in severe clinical populations, a test developer might choose items that provide maximum information within the higher levels of theta. On the other hand if choosing items for a

screening measure to be used with general population, a set of items with maximum information at a variety of levels of theta might be selected. In order to compute the item information curve, which is considered continuous, information values are computed at each level of theta using the discrimination parameter and difficulty parameter(s).

According to Embretson and Reise (2000) category response curves can be transformed into item information curves with the following formula.

$$I_j(\theta) = \sum_{k=0}^{m} \frac{P_{jk}^{*}(\theta)^2}{P_{jk}(\theta)} \qquad (6)$$

The total scale or test information function is computed by summing the individual item information curves.

Sample Size and the GRM.

Lautenschlager, Meade, and Kim (2006) examined the performance of the GRM under varying conditions that included different sample sizes and test length. Using data from 891 respondents to the Minnesota Satisfaction Questionnaire they computed item parameters, which they then used to generate simulated respondent data sets. They created data sets with 5, 10, 15, or 20 items with sample sizes of 75, 150, 200, 300, 500, 1000, and 2000. Each data set was then analyzed with Multilog to compute item parameters, these parameters were then linked to the original parameters, and mean squared errors were computed for both item and person parameters (α, β, & θ). These estimates of MSE were compared to the values from the original data set. As well, correlations between the original and generated parameters were compared. The results indicated that sample sizes of 75 and 150 resulted in significant sampling errors with respect to item parameters that were largely unaffected by test length. With respect to

estimates of θ it was found that, although test length did affect the stability of the estimates, with shorter scales having a deleterious effect, correlations with the original estimates were strong given a sample size of 300 or larger. These authors concluded that a small sample size and limited numbers of items reduce the accuracy of item parameters and recommend a sample size of at least 300 when working with short scales (e.g., five items). Reise and Yu (1990) examined the ability of the software Multilog (marginal maximum likelihood estimation) to recover estimations of parameter in the GRM. These authors concluded that a minimum sample size of 500 respondents is needed to attain adequate structural parameters while adequate person parameters can be attained with 250 to 500 respondents.

Equating Parameters, Differential Item Functioning and Differential Test Functioning

Differential Functioning of Items and Test

According to Hidalgo-Montesinos and Lopez-Pina (2002) differential item functioning occurs when item response functions vary across different groups when matched upon the underlying trait. These authors indicate that identifying and removing items demonstrating DIF reduces the probability of false positives. Alternatively, DIF has been said to occur when the probability of endorsing a category or option is influenced not only the individual's standing on the latent trait intended to be measures but is also influenced by the respondent's membership in a population (Morales, Flowers, Gutierrez, Kleinman, & Teresi). The presence of DIF may be due to systematic differences in the way in which one group interprets the questions or options. Collins, Raju, and Edwards (2000) state that a scale can indicate differences between groups because of actual differences or because the scale does not accurately measure the trait in one or more of

the comparison groups. If a scale indicates difference between groups because of the latter possibility, the results of using the scale in such populations may be invalid and misleading. The assessment of DIF using IRT has become an increasing popular method often based upon sex, cultural group, and age.

In 1995, Raju, van der Linden, and Fleer developed a parametric item response theory based procedure for assessing differential item functioning (DIF) and differential test functioning (DTF). They indicated that this framework, entitled Differential Functioning of Items and Tests (DFIT) could be employed in the examination of either unidimensional or multidimensional data score dichotomously or polytomously.

According to Flowers et al. (1999), the DFIT framework is unique in several respects. It is the only parametric IRT method that assesses differential functioning at the level of the items and the test. While examining an item this framework does not assume that all other items are unbiased and provides the test developer with the opportunity to examine the overall effects of removing an item from a test. It also allows the flexibility to examine tests that are a combination of both dichotomously or polytomously scored items. This group of researchers has offered empirical data to support the use of this framework in the examination of dichotomous data, dichotomous unidimensional data, and polytomous unidimensional data (Raju et al., 1995; Flowers, Oshima, & Raju, 1999).

Flowers et al. (1999) described the use of the DFIT framework in the examination of polytomous unidimensional data and explained the manner in which a researcher might detect DIF and DTF with Samejima's (1969) graded response model. In this model the calculation of boundary response functions (BRF) and item category response functions (ICRF) is completed in order to estimate the probability of response in each

category. From these probabilities it is possible to calculate expected item scores (ES), plot expected item score functions (IRF), and expected test score (ETS) functions for each subject. An item is considered to be exhibiting differential functioning if the ES for an individual in the reference group with a given θ, scores in a different category than an individual in the focal group with the same θ. Similarly, a test is considered to be functioning differentially if the ETS for both groups with the same θ are not equal.

In order to complete these analyses within the DFIT framework it is necessary to estimate item parameters for the reference group and the focal group, which are linked onto the metric of the focal group parameters via a linear transformation. Two expected scores are then calculated for each member of the focal group for each item and the test. For each item and the test, a score is calculated using the focal group parameters and another is calculated employing the linked reference group parameters. If the scores are not equal, the test or item is thought to be functioning differentially.

An important and unique feature of the DFIT framework is the estimation of two different forms of DIF; compensatory differential item functioning (CDIF) and non compensatory (NCDIF). CDIF reflects the additive or directional nature of DIF, in that one item from a scale may favour one group while another item from the same scale may favour the other. The net effect would be a cancellation of each other. CDIF is calculated by summing the DIF for each item on a scale and provides directionality. Raju et al. (1995) indicate that CDIF can be calculated with the following formula.

$$\text{CDIF}_i = \in (d_i D) = \text{Cov}(d_i, D) + \mu_{d_i}\mu_D, \qquad (7)$$

Where $d_{is} = P_{iF}(\theta_s) - P_{iR}(\theta_s)$, D_s equals T_{sF} - T_{sR}, s refers to examinees, and T equals true score or expected proportion correct. According to Raju et al. (1995), DTF can be calculated by summing of the CDIF for each item as follows.

$$\text{DTF} = \sum_{i=1}^{n} \text{CDIF}_i , \qquad (8)$$

Where n is the number of items.

Flower et al. further suggest that test makers could use CDIF values and directions to determine which items to keep on a test or which items to eliminate. In this manner the bias from one item may be compensated for or balanced by the bias in another item. NCDIF is an estimate of an item's DIF, under the assumption that all other items on the test are free of DIF and is therefore more similar to other IRT measures of DIF. All values for NCDIF are positive and therefore cannot be viewed as compensatory. According to Raju et al. (1995) can be calculated employing the following formula.

$$NCDIF_i = \sigma_{d_i}^2 + \mu_{d_i}^2 , \qquad (9)$$

Where σ_{d_i} represent the variance of d_i and μ_{d_i} represents the mean.

Collins, Raju, and Edwards (2000) employed the DFIT framework to examine a 10 item scale intended to measure an individual's work satisfaction which employed a

five option Likert-type scale. The purpose of this study was to assess differential functioning using the DFIT framework as compared to two other methods; Lord's chi square and the extended signed area. In this example the test was compared when used in a sample of Caucasians versus African Americans and males versus females. The analysis in this study began with a factor analysis using SPSS to confirm unidimensionality of the scale. This procedure was followed by the estimation of item parameters (A & K) using Parscale, the equating of the parameters to a common metric with EQUATE 2.0, and then determining the presence of DIF and or DTF utilizing the three previously discussed methods. In order to compare the results of the three different methods, the authors examined the results of one item from the scale in detail. This item was found to exhibit DIF regardless of the method chosen. Significant differential functioning was not found at the test level in either of the comparisons. The authors concluded that the DFIT framework produced more consistent results than either the Lord chi square or the signed area methods.

Practical Application of DFIT.

According to Raju, van der Linden, and Fleer (1995) there are several practical applications of the DFIT framework and its associated measures; CDIF, NCDIF, and DTF. As is often the case, there is no one algorithm to determine which statistic to employ that will be appropriate in all test construction situations. Raju et al. suggest that when the total test score is of major interest that DTF is the most appropriate measure to use to determine items to include. In most situations it will be necessary to retain items that display some bias and in such situations CDIF is considered the most useful measure (Raju et al., 1995) to aid in balancing a measure. NCDIF is most useful in determining

items that exhibit the highest levels of bias towards one group. In the case of this dissertation, all three of these statistics will be considered depending upon the decision making context.

DIF analysis from an IRT perspective has been employed in more than one instance to examine the utility of the Center for Epidemiologic Studies Depression Scale (CES-D) in different cultural and demographic populations. For example, Iwata, Turner, and Lloyd (2002) conducted a DIF analysis to examine the manifestation of symptoms of depression in young adult African-Americans, Hispanics (born in the US), Hispanics (born outside the US) and non-Hispanic Caucasians. They found that approximately half of the items of the CES-D functioned differentially between the non-Hispanic Caucasian group and the other three cultural groups with significant Spearman partial correlation coefficients ranging from .078 to .215 (p values of less than .017 or smaller). They also found that African-Americans tended over-endorse somatic symptoms (Spearman partial correlation = 0.104, $p < .002$), while under-endorsing depressed affect (Spearman partial correlation = -0.165, $p < .002$) as compared to white non-Hispanic Americans, suggesting cultural difference in expression of distress. This finding demonstrates that IRT analyses can serve to empirically delineate or clarify the essential symptoms of a phenomenon in a specific sample.

Purpose of This Research

The purpose of this dissertation is the examination of the psychometric function of the PAI from an IRT perspective with a sample of methadone maintenance treatment patients resulting in an improved scoring algorithm. It is hypothesized that this effort will also demonstrate the ability of IRT methods to eliminate the least effective items from a

scale without sacrificing the majority of the information offered by the scale. This research will employ a variety of nonparametric and parametric IRT methods designed to examine the psychometric function of the scales and subscales of the PAI within a sample of methadone maintenance treatment patients as compared to a post-secondary student population. The items and scales will be examined for monotonicity, unidimensionality, information, and differential functioning. It is hypothesized that a new more effective scoring system will be developed through the elimination of items that do not conform to the assumptions of IRT models, offer the least information, or display significant levels of differential item functioning that result in significant differential test functioning. These analyses will include a series of IRT examination of each PAI scale and subscale with the exception of the INF validity scale which does not conform to the Graded Response Model.

For example, a new version of the BASIS-32 dubbed the BASIS-R was recently developed using a combination of factor analysis, Classic Test Theory, and Item Response Theory. Not only were 8 of the original 32 items eliminated, but the developers were able to create new more psychometrically sound scales (Eisen, Normand, Belanger, Spiro, & Esch. 2004).

Method

Subjects

Data for this dissertation were combined from several archival sources of de-identified records. Included for comparison were a large sample of university students (919) and a large diverse sample of individuals receiving methadone maintenance treatment (323). The students (73% female; mean age of 22.3, standard deviation of 5.8)

volunteered to participate in return for course credit in an Introductory Psychology course at a small rural Canadian University. The clinical sample consisted of the combination of two separate archival data sets. One sample was comprised of 239 inner city methadone patients (39% female) consisting of mainly African American (46%) and Latino (49%) individuals from a large U.S. metropolitan area. On average these individuals had completed 11.5 years of education with an average age of 37 years. The second clinical sample consisted of 95 (47 male, 48 female) individuals from a small Canadian city. The mean age of this sample was 33.38 years with a standard deviation of 9.14. 80% of the sample was comprised of Caucasian Canadians and 20% Aboriginal Canadians. On average these individuals had attended school for 11.1 years with a standard deviation of 2.34.

The PAI administrations from the comparison group were screened for individuals with a possible substance use disorder based upon their scores on the DRG scale of the PAI. Subjects were disqualified from the comparison groups based upon the cutoff scores suggested by the authors of the PAI. The test authors suggest that a T Score on the DRG subscale over 70 is consistent with substance abuse. A score over 80 is consistent with substance dependence. Based on these cutoffs, any individual in the comparison groups receiving a T Score on the DRG score over 70 was eliminated from the analysis.

Analyses

The analyses for this dissertation consisted of a series of NIRT and PIRT methods employing several different models and software packages. The employment of more than one IRT method including both NIRT and PIRT analyses is often recommended (de

Konig, Sijtsma, & Hammers, 2002) as the initial examination of the data employing a NIRT model allows the researcher to examine the structure of the data to determine if the structure matches the PIRT model to be employed. Prior to the initiation of IRT analyses, all reverse scored items of the PAI were recoded and data files constructed for the various analsyes. The NIRT analysis first involved the plotting of option and item characteristic curves employing Testgraf in order to examine the data visually. The purpose of plotting option and item characteristic curves was to examine the scales and subscales with respect to the assumptions of monotonicity and to determine the appropriateness of the GRM with respect to the assumption of a hierarchical progression through the item options. This plotting of curves was followed by Mokken scale analysis procedures employing MSP5. The purpose of Mokken scale analysis was to determine the presence of unidimensional clusters of items that display monotone homogeneity. Items which clearly do not conform to these assumptions were dropped from further consideration. The next phase of the analysis consisted of the calibration of item discrimination and difficulty parameters using the parametric Graded Response Model employing the Multilog software. Again any items which did not meet the minimum cut offs were eliminated. All remaining potential items for a scale or subscale were then examined for DIF or DTF through the DFIT framework (which entail two software programs: EQUATE and DFITP5). Items which contributed to significant DTF were immediately dropped from the analysis, while significant NCDIF was a factor considered in the final item retention process. The final item retention algorithm determined the final items retained. In this final phase, the results of all the analyses were integrated and considered in an attempt to retain only the most effective items offering the maximal information.

Once the final items were selected for each revised scale or subscale, option characteristic curves, item characteristic curves, item parameters, scalability coefficients, item information curves, and total scale information curves were re-calibrated and plotted to verify improved functioning of the scales. Alpha coefficients were computed to compare the internal consistency reliability of the revised scales and subscales with that of the original scales and subscales in this and other populations from previous studies.

Results

Aggression Subscales

Aggressive Attitude (AGG- A)

The functioning of the items of the AGG-A subscale of the PAI was initially examined through the computing of option and item characteristic curves using Testgraf (Figure 3.). Item #258 (bad temper) displayed some minor violations of monotonicity and of the GRM in that option category #3 was always more likely to be chosen than option category #2 across all values of theta. Item #259 (hard to anger) displayed similar violations of monotonicity. The other four items of the scale appeared to conform to the assumption of monotonicity and the options appeared to be functioning as expected by the GRM.

The monotone homogeneity of the PAI AGG-A subscale was examined through Mokken Scaling analysis. Based upon the guidelines suggested for interpretation the AGG-A subscale demonstrated a poor fit with the Mokken Model (scale $H = .29$). Items #259 and 339 demonstrated item H values below .30. Given this indicator of weak scalability, the SEARCH function of the MSP5 software was employed to determine if a dominant cluster of items existed within the subscale. This analysis revealed a scale with

a low level of fit and acceptable reliability (scale H = .35, Rho = .75). As can be seen in Table 3, this scale contains five of the original 6 items and eliminates item #259 from the subscale.

Item parameters were calibrated for the 5 remaining items of the PAI subscale AGG-A (Table 3). Items #298 and #299 were found to have discriminatory parameters below 1 and widely divergent difficulty parameters. These two items were dropped from the analysis and the remaining three items were again examined and found to have acceptable item parameters as displayed in Table 3. The remaining three items of the PAI AGG-A-R subscale were examined for the presence of DIF and or DTF. In this case no significant DIF or DTF was discovered.

As a result of this nonparametric and parametric IRT analysis a three item revised subscale was derived that included items #258, #338, and #339. These three items demonstrated a medium fit with the Mokken model (scale H = .48), acceptable reliability (Rho = .72, α = .71), and acceptable item parameters as calibrated by the Multilog software (Table 3). OCC and ICC curves were re-calibrated for the three retained items and indicated a good fit with the GRM model, good discrimination across the continuum, and strong monotonicity. Item information curves for the final three remaining items are displayed in Figure 4. Items #258 (bad temper) and 339 (hard to calm down) consistently demonstrated the highest levels of information on this scale intended to assess aggressive attitude. Figure 5 compares the information offered by the original scale and that of the revised version. As can be seen, the elimination of 50% of the items resulted in minimal information loss.

Physical Aggression (AGG-P)

Item and option characteristic curves were plotted for the six items of the original AGG-P subscale of the PAI and are displayed in Figure 6. All items of this subscale appeared to conform to the assumption of monotonicity. Visual inspection of these curves indicated that Item # 221 failed to demonstrate good discriminative properties above a theta value of -1. The majority of items displayed some inconsistencies with the GRM. For example, in the case of Item #61 option #1 is always a more probable response than option #2 across the entire range of theta.

A Mokken scale analysis confirmed that Item #221 did not conform to the assumption of monotonicity. This item will not be included in the set of items examined for homogeneity. The results indicate that these remaining five items do form a scale that displays a moderate fit with the Mokken model (scale $H = .49$). Individual item scalability coefficients are displayed in Table 4

Item parameters were calibrated for the 5 remaining items of the PAI subscale AGG-P (Table 4). All item parameters were found to be acceptable. These remaining five items of the AGG-P scale were analyzed for DIF and DTF employing the DFIT Framework. Significant levels of DIF and or DTF were not indicated.

Given that the remaining five items of the AGG-P subscale all conform to the assumptions of monotonicity and unidimensionality and do not demonstrate DIF or DTF, item retention was based upon the item discrimination and difficulty parameters. Therefore, items #21, #61, and #101 were retained for the AGG-P-R subscale for use in MMT populations. The final scale was re-analyzed to examine the level of scalability; a strong fit with the model (scale $H = .60$, $\alpha = .67$) was indicated. Item parameters and scalability coefficients for the final scale are shown in Table 4. OCC and ICC curves

were recalibrated for the three retained items, indicating a good fit with the GRM model, good discrimination across the continuum, and strong monotonicity.

Item information curves are displayed in Figure 7 , clearly demonstrating that Item #61 (temper explodes, lose control) offers the most information from the available items when employed in this population. Figure 8 demonstrated a minimal loss of overall information given the reduction of 50% of the items.

Verbal Aggression (AGG-V)

Item characteristic curves and the option characteristic curves were plotted for all items of the AGG-V subscale of the PAI when administered to the MMT sample and are displayed in Figure 9. While visual inspection did not reveal obvious violations of monotonicity, they did raise questions as to whether they fit the graded response model.

Mokken scaling analysis indicated that the AGG-V subscale had a poor fit with the Mokken model (scale $H = .12$), with no individual items meeting the minimum H value of .30. Therefore, all items of the AGG-V subscale were examined employing the SEARCH function of MSP5 to determine if any clusters existed. Two separate factors were discovered, each with two items. The first cluster involved items #58 and #138. Together these items displayed the strongest relationship (scale $H = .52$). When examining the items they do seem to possess strong face validity in that they both involve endorsing aggressive verbal behaviour towards others. The second weaker cluster is composed of items #178 and #218 (scale $H = .32$) and involve items that ask the respondent to endorse the avoidance and the dislike of raising his or her voice or arguing. Given the fact that this H value is barely acceptable it was decided to retain only the first cluster for the AGG-V-R subscale. Item parameters will be estimated for these two items

followed by an investigation employing the DFIT framework. Of note, initial item parameter estimation of the original six items would have also resulted in the retention of the two same items.

Item parameters were calibrated for this two item scale and were found to be acceptable ($\alpha = .79$). Next the two items were examined for DIF and DTF employing the DFIT Framework. Significant levels of DIF and or DTF were not indicated.

Items #58 and #138 were the only two items deemed to meet inclusion criteria. The item parameters and the scalability coefficients for this scale are displayed in Table 5. As demonstrated in Figure 10, these two retained items offer adequate levels of information within the middle of the continuum, while Figure 11 demonstrates that the 2 item AGG-V-R scale offers similar information as the original 6 item scale. OCC and ICC curves were computed for the revised scale indicating much improved fit with the GRM, discrimination across the continuum, and monotonicity.

Alcohol Scale

Item characteristic curves and option characteristic curves were plotted for all items of the ALC scale of the PAI when administered to the MMT sample and are displayed in

Figure 12. Visual inspection of these plots indicated that items #294, 295, and 334 violated the assumption of monotonicity. Most items displayed some irregularities with respect to the hierarchical ordering of option choices as assumed by the GRM.

Initial examination of the ALC scale for fit with the Mokken model using the Test function of MSP5 indicated a moderate level of scalability. (scale $H = .43$, Rho = .87). As

can be seen in Table 5, Item #294 displayed an unacceptable scalability coefficient and was therefore removed from the list of potential items prior to the PIRT analysis.

Item parameters were calculated for the remaining 11 items of the ALC scale and are reported in Table 7 Items #295 and #334 demonstrated insufficient discriminatory power and were eliminated from further analysis. The item parameters for the remaining 9 items were re-calibrated and are presented in Table 8 along with their re-calibrated scalability coefficients. An analysis of differential item and test functioning was conducting employing DFIT software based upon item parameters obtained from Multilog and Equate software. Results indicated significant levels of NCDIF on item #175 (Table 8) however no significant DTF was indicated.

Final item selection for the ALC-R scale will be based upon the various parameters estimated in this analysis. Items #294 was eliminated as it violated the assumptions of monotonicity and homogeneity. Items #295 and #334 were eliminated due to low discriminatory power, while Item #175 was eliminated due to the presence of differential item functioning. At this point eight items remained that all displayed acceptable characteristics. Items #254 and #335 were dropped as they offered the least information amongst the remaining items. The final set of items for the revised ALC-R consisted of items #15, #55, #95, #135, #215, and #255. These items were analyzed one final time to arrive at estimates of item parameters and scalability coefficients (Table 9). The final scale demonstrated a high level of fit with the Mokken Model (scale $H = .62$, Rho = .89, $\alpha = .88$). This suggests that individuals can be ordered in the severity of their alcohol use based upon the scores they receive on this scale. Item information curves are displayed in Figure 13 while Figure 14 compares the total information provided by the

original 12 items to the final six. As can be seen the elimination of six items does not substantially reduce the total information offered. The content of the final scale includes the knowledge one needs to reduce use, difficulty in controlling use, being told one needs to reduce use, feeling guilty over use, and experiencing problems at home and in relationships due to alcohol use. Item #135, which asks the respondent to endorse relationship problems due to his or her alcohol use, was found to provide the most information or precision.

Antisocial Subscales

Antisocial Behaviours (ANT-A)

Option characteristic curves (OCC) and item characteristic curves (ICC) were plotted for each item of the ANT A scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 15). Overall, the majority of items were shown to be monotonic when employed in this population. Item 51 did display a mild level of violation of this assumption at the extreme end of the range of θ. The option characteristic curves tended to suggest that the GRM assumption of ordered options is being violated in some instances. .

Analysis of the full scale with the Mokken Model employing the MSP5 software indicates the presence of three factors within this one subscale. The first factor consists of items #51, #91, #251, and #291. The content of these items reflected criminal behaviours. Two other factors were revealed, the first of which consisted of items #11 and #211 concerning truancy and behaviour problems at school. A third factor consisted of items #131 and #171 which both describe lying and escaping tight situations. This presents a difficult decision in that all three of these factors represent areas of DSM-IV-TR criteria

for ASPD (or Conduct Disorder which is a required precursor), yet as a scale they seem to lack significant relation to each other. Since factor one seems to be the factor most clearly connected to the DSM-IV-TR criteria, it was decided that first factor would be retained for further consideration. Estimation of item parameters of these four items with Multilog resulted in a solution that did not satisfy the intercycle parameter change maximum value. For that reason a further examination of the dimensionality was conducted. A different algorithm was employed in which items were removed based upon the lowest scalability score. Using this method a new four item scale was developed (scale $H = .36$, Rho = .60, $\alpha = .60$). This new item arrangement included two items querying criminal type behaviour and two item querying lying type behaviours. This new scale was examined with Multilog for the purposes of determining item parameters and determining if this new configuration would better fit the GSM.

The calculation of item parameters was conducted and these results are displayed in. displays the items retained from the original ANT-A scale for use in the assessment of MMT populations including item parameters and scalability coefficients. Overall it was found that the most effective items in this scale focused upon illegal activities, property damage, lying, and avoiding detection. Overall, the least effective items asked the individual to endorse problems or expulsion from school, trouble with the law, and stealing. Information curves for the retained items of the ANT-A-R subscale are displayed in Figure 16, while total subscale information curves comparing the new subscale to the old can be viewed in Figure 17. Minimal information loss is evident; however, the area of maximum information offered has shifted appropriately to the more severe end of the continuum.

Egocentricity (ANT-E)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANT E scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 18). Overall, the majority of items were shown to be monotonic when employed in this population. Items 5 and 6 did display a mild level of violation of this assumption at the extreme end of the range of θ. However, the option characteristic curves of most of these items displayed some irregularities with respect to the GRM assumption of ordered response categories.

This concern was verified by the Mokken analysis where the results again suggested a lack of monotonicity and the presence of multiple dimensions (scale $H = .24$, Rho = .68). The scale was then examined employing the SEARCH item selection method to determine the existence of item clusters. Three clusters were identified. Each of the three clusters was examined with Multilog and item parameters were estimated in order to determine which of the clusters was to be retained.

Each cluster identified in the Mokken analysis was examined with Multilog. The first cluster consisted of Items #71, #111, and #151 dealing with behaving in a manner taking advantage of others for personal gain. The initial analysis indicated that the scale did not conform to the model and that in particular Item #111 displayed unacceptable item parameters. The analysis was conducted a second time following the removal of that item. This analysis revealed that the data now better fit the model and that both items displayed adequate discriminatory and difficulty parameters (Table 11).

The second cluster was also analyzed for model fit and item parameters. This cluster consisted of two items (#271 and #211) that deal with egocentric attitude. Satisfactory fit and acceptable item parameters (Table 11) were found.

The final cluster consisting of items #191 and #231 was examined with Multilog. It was found that these two items had unacceptable item parameters and they were dropped from consideration. Both potential subscales were examined employing the DFIT framework were found to be free of both DIF and DTF.

It was decided that only the first cluster would be retained given its moderate scalability coefficient and weak reliability ($\alpha = .48$). Figure 19 displays the item information curves for these two items.

Stimulus Seeking (ANT-S)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANT S scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 20). All items on this scale, when employed in the MMT population, conformed to the assumption of monotonicity.

However examination of the scale with MSP5 indicated a poor fit with the Mokken model (scale $H = .25$, Rho = .68). Implementation of the SEARCH item selection process indicated one central scale consisting of items #39, #79, #119, 159, and #279, which seem to examine issues around finding pleasure in dangerous activities (scale $H = .39$, Rho = .72). Item scalability coefficients are displayed in Table 11.

The five items determined to constitute a scale as a result of the nonparametric analysis were examined and item parameters calibrated using the Multilog software. This analysis indicated that item #159 had a discriminatory parameter below one, which

eliminates this item from the scale (

Table 12).

The scalability coefficient and item parameters were estimated for the final four items of the proposed subscale ANT-S-R (

Table 12). Item and subscale information curves are presented in Figure 21 and Figure 22. Minimal information loss is observed in the removal of four items. Item #79 which asks the respondent to endorse engaging in wild behaviours for fun was demonstrated to offer the most information of the retained items. This revised scale demonstrated acceptable reliability ($\alpha = .75$).

Anxiety Subscales

Affective (ANX-A)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANX A scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 23). Item #244 displayed serious violation of the assumption of monotonicity, while items #124 and #164 displayed mild violations. The items tended to meet the assumptions of the GRM with respect to ordered response options.

These results were verified by an analysis using MSP5 which indicated that the scale demonstrated a weak fit with the Mokken mode. (scale $H = .26$, Rho = .71). A further analysis employing the MPS5 SEARCH function indicated a 5 item facet within the PAI ANX-A subscale (Table 13). Therefore these five items will continue in the analysis and will have item parameters calibrated.

Item parameters were estimated for the five remaining items of the ANX-A scale (Table 13). All discrimination and difficulty parameters were found to be within acceptable parameters. Therefore these five remaining items were examined for DIF and DTF employing the DFIT framework. Item #4 and #284 exhibited significant levels of DIF (Table 13), however no significant DTF was observed.

Of the five remaining items, #284 was chosen to be eliminated since it displayed significant levels of DIF and also offered the least amount of information. The four remaining items #4, #44, #84, and #204 were re-examined to determine the final item parameters (Table 13), estimates of available information (Figure 24 and Figure 25), and scalability coefficients (Table 13). Item #44 which asked the respondent to endorse experiencing nervousness was demonstrated to offer the most information of the retained items. Option characteristic curves and item characteristic curves were nonparametrically re-plotted using Gaussian kernel averaging. The resultant plots indicated an improved fit with the graded response assumptions and monotonicity. This revised scale demonstrated acceptable reliability ($\alpha = .76$).

Cognitive (ANX-C)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANX C scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 26). Items #185 and #225 displayed serious violations of monotonicity in the middle of the distribution of the underlying trait, while #305 displayed moderate violation of the assumption of monotonicity in the more severe region of theta. Items #185 and #225 displayed non zero lower asymptotes, indicating that even those respondents with the lowest score on the scale were likely to endorse mild symptomatology.

An initial examination of the ANX-C scale based upon responses from a sample of MMT patients indicated a poor fit with the Mokken model (scale $H = .28$, Rho = .73). The results confirmed that items #185 and #225 do not conform to the assumption on monotone homogeneity and that item #8 may not conform. Based upon these results a

second analysis was conducted using the SEARCH function to determine the presence of a central factor. This analysis revealed one dominant cluster made up of items #25, 65,105,145, 265, and 305. This cluster displayed a moderate fit with the Mokken model (scale H = .52, Rho = .84). Item #185 and #225 were also found to form a weak cluster (scale H = .32, Rho = .49) based upon item content asking the respondent to endorse an absence of worrying behaviours. It was decided based upon the superior model fit, to continue the analysis with the first cluster.

Initial item parameter estimation was conducted using Multilog with the remaining 6 items of the ANX-C subscale in the MMT population. As seen in

Table 14, all items displayed acceptable parameters and were included in the next phase of the analysis to determine if DIF or DTF is present in the scale. A comparison of the functioning of the remaining six items of the ANX-C scale in the MMT population and the student population revealed no significant levels of DIF or DTF.

The final items selected for the new ANX-C-R subscale were chosen based upon item information within the MMT population. Items #25, #65, #105, and #265 were retained and item parameters and scalability coefficients are displayed in

Table 14. The new scale demonstrated a moderate fit with the model (scale H = .58, Rho = .82, α = 81). Item information curves are demonstrated in Figure 27. As demonstrated Item #105 (so worried, can't stand it) offered the most information of any of the retained items. Figure 28 compares the information curves based upon the original items compared to the total information of the reduced ANX-C-R. As can be seen there is minimal information loss incurred by the removal of half of the items of the subscale.

Option characteristic curves and item characteristic curves were nonparametrically re-plotted using Gaussian kernel averaging. The resultant plots indicated an improved fit with the graded response assumptions and monotonicity.

Physiological (ANX-P)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANX P scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 29). Visual inspection of the plots indicated mild violations of monotonicity involving items #193 and #313 in the middle of the theta distribution. Items #113, 153, and 233 exhibited minor violation of the assumption of ordered response categories of the GRM.

Analysis of the ANX-P subscale indicated a poor fit with the Mokken Model (scale H = .30, Rho = .75). This analysis also found items #193 and 313 to not conform to the assumptions of monotone homogeneity. These results indicated the need to examine the scale for the presence of a central facet using the SEARCH function of MSP5. This examination resulted in a six item scale that demonstrated a moderate fit with the Mokken model (scale H = .42, Rho = .79). These items were retained for parametric analysis.

As demonstrated in Table 15 all item parameters for the remaining six items of the ANX-P subscale were within acceptable ranges, therefore these six items will be examined for DIF and or DTF. The remaining items were examined employing the DFIT framework and were found to exhibit no significant DIF or DTF. Therefore item retention will be base upon maximizing the available information.

Based upon item parameters it was decided to retain #73, #113, #153, and #233 were retained based upon maximizing information. Item parameters and scalability coefficients for the four items are displayed in Table 15. The scale displayed a medium fit with the Mokken Model (scale H = .45, Rho = .73, α = .73). Figure 30 demonstrates the item information curves for the final set of ANX-P-R which indicates that item #113 (dizzy under pressure) offers the most information of the retained items. Figure 31 compares the total information of the new 4 item subscale versus the original 8 item subscale.

Anxiety Related Disorders Subscales

Obsessive-Compulsive (ARD-O)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ARD O scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 32). Item #125 displayed serious violation of the assumption of monotonicity, while items #5 and #286 displayed mild violations.

An initial analysis of the scale confirmed the results discussed above. The scale had a poor fit with the Mokken Model (scale H = .18, Rho = .62), while Item #125 exhibited a negative item scalability coefficient. Implementation of the Search function of

MSP5 indicated the presence of two factors. The first factor included items #5, # 45, and # 85 that appears to reflect symptoms and behaviours associated with OCD (Table 16). This scale demonstrated a strong fit to the Mokken Model (scale H = .50, Rho = .71). The second cluster includes two items that seem to reflect symptoms of Obsessive Compulsive Personality Disorder that query perfectionism and attention to detail (Table 23), and displayed a weaker fit with the Mokken Model and weaker reliability (scale H = .35, Rho = .50). It was decided to retain the first cluster since it demonstrated acceptable monotone homogeneity and reliability.

The graded response model was calibrated to the retained items and the item parameters are reported in Table 16. Acceptable parameters were found for all items. Analysis of this subscale did not reveal significant DIF or DTF. This revised scale demonstrated weak reliability (α = .64).

OCC and ICC were re-plotted for the new ARD-O-R subscale containing three items. These plots indicated that the items now conformed to the GRM, were monotonic, and discriminated well across the continuum. Figure 33 demonstrates item information curves for the ARD-OCD-R. Item #5 which asked the respondent to endorse the need to perform actions in a certain manner was demonstrated to be the most discriminating item.

Phobias (ARD-P)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ARD P scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 34). Items #165, 245, and 285 displayed serious violation of the assumption of monotonicity. Items #106, 266, and

306 did not display the hierarchical ordering of option categories expected under the GRM.

Mokken scale analysis indicated that this scale did not display monotone homogeneity (scale $H = .18$, Rho = .62). The scale was then examined to determine if a central factor existed. This analysis indicated one central cluster that consisted of items #26, #66, and #106. Examination of the content of these items indicated that this cluster represented a social phobia factor. A second factor was demonstrated that consisted of two items; #186 and #226 which both suggest content around simple phobias. The first factor displayed superior *H* coefficients and reliability. For that reason, it was decided that this scale would be reduced to three items that appear to be measuring aspects of social phobia. These three items would be the basis of the following analysis.

An initial examination of all the original items of this scale employing Multilog also indicated that only these three items displayed acceptable parameters and conformed to the assumptions of the GRM. Significant levels of DIF or DTF were not indicated in this analysis.

The three items of the ARD-P-R subscale were examined in a PIRT analysis with the GRM employing Multilog. The item parameters were found to be quite acceptable and are displayed in Table 17 along with the item scalability coefficients. This revised scale demonstrated acceptable reliability ($\alpha = .70$). Item information curves of the new subscale are display in Figure 35. Item #66, which asked the respondent to endorse experiencing exaggerated fears, was found to offer the most information amongst the three remaining items. Figure 36 compares the information of the new reduced item scale with that of the original larger scale, demonstrating that little information is lost

through the reduction of 50% of the items. Re-plotting of the OCC and the ICC indicated that this new scale conformed to the GRM, was monotonic, and discriminated across the entire range of theta.

Traumatic Stress (ARD-T)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANX P scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 37). Visual inspection of these curves did not indicate any violations of monotonicity. Minor violations of the expected hierarchical ordering of response categories were noted.

This was supported by the results from an initial analysis with MSP5 which indicated that all items displayed monotonicity. This analysis also supported the presence of one major factor within this scale. Therefore all items will be retained for analysis with Multilog and the estimation of item parameters.

As seen in Table 18, item parameters were calibrated for all items of the ARD-T subscale. All items were found to exhibit sufficient discrimination and difficulty parameters to be included in the next stage of the analysis to determine if DIF or DTF is present. The eight items of the original PAI subscale ARD-T did not exhibit significant DIF or DTF. Therefore the final items will be chosen based upon maximizing information.

Since all items of the scale conform to the required assumptions and in addition no significant DIF or DTF was indicated, the items were chosen in order to maximize the total information provided by the scale. Items #34, #114, #154, and #234 were chosen and analyzed for item parameters and scalability coefficients (Table 19).

Re-calibration of option and item characteristic curves indicated the scale now conformed to the hierarchical ordering of response categories assumed by the GRM. The scale showed a strong level of fit with the Mokken Model (scale H = .67, Rho = .87, α = .70).

Figure 38 displays the item information curves. Item #114, which asked respondents to endorse experiencing memories of a bad experience was found to offer the most information and to be the most discriminating item of the scale. Figure 39 compares the total information offered by the ARD-T-T subscale (4 items) as compared to the original 8 item scale indicating that the scale offers maximum levels of information in the higher ranges of the underlying trait.

Borderline Features Subscales

Affective Instability (BOR-A)

Option and item characteristic curves were plotted using Gaussian kernel smoothing techniques (Figure 40). Item # 134 is observed to violate the assumption of monotonicity. Items #14, 134, and 214 displayed minor violations of the GRM assumptions regarding the hierarchical order of response categories.

The BOR-A scale was examined and found to exhibit a poor fit with the Mokken Model (scale H = .25, Rho = .60). The Search function of the MSP5 software was employed in order to determine the presence of a central facet or cluster. A central facet was discovered consisting of three items consisting of #14, #54, and # 214 that demonstrated a moderate fit with the Mokken model and acceptable reliability (scale H = .50, Rho = .72). Item parameters were then calibrated for these items.

Item parameters were estimated for the three remaining items of the PAI subscale BOR-A (Table 20). All parameters were within recommended acceptable parameters; therefore these items were examined for differential functioning. No significant DIF or DTF was demonstrated in the three remaining items. This revised scale demonstrated acceptable reliability ($\alpha = .72$).

The BOR-A-R consists of the previously discussed three items. Option and item characteristic curves were re-plotted employing a Gaussian kernel smoothing techniques and much improvement was found with respect to monotonicity and model fit. Item information curves were plotted (Figure 41) demonstrating that item #54 (intense mood shifts) offered the maximum information of these remaining items. Figure 42 compares the total information provided by the original six item BOR-A subscale as compared to the three item BOR-A-R subscale indicating little loss of information.

Identity Problems (BOR-I)

Individual option and item characteristic curves were plotted for the six items of the BOR-I (Figure 43). It was found that item #217 violates the assumption of monotonicity. Item #57 displayed a minor violation of monotonicity. All items conformed to the assumption of hierarchical ordered response categories.

The six items of the BOR-I subscale displayed a weak fit with the Mokken model (scale $H = 0.31$, Rho = .71). Given this result, the Search function of MSP5 was employed to determine the presence of a central facet. One central facet consisting of five items was demonstrated with a moderate fit with the Mokken model (scale $H = 0.46$, Rho = 0.79, $\alpha = .76$). The results indicate that item #217 (not easily bored) should not be maintained in the next level of analysis.

Discrimination and difficulty parameters were calibrated using the GRM for the remaining items (Table 21). All parameters were found to meet the minimum requirements; therefore these five items were examined for DIF and DTF. Item parameters were equated between the two groups and then examined for DIF and DTF. Items #57 and 177 displayed significant DIF; however significant DTF was not demonstrated (

Table 21).

Based upon the various analyses and the lack of DIF, item retention was based upon information. For that reason, items that offer the most information were chosen resulting in a final scale consisting of item #17, 57, and 114. Item parameters and scalability coefficients are displayed in

Table 21. The scale demonstrated a strong fit with the Mokken model and good reliability (scale $H = 0.55$, Rho = 0.77, $\alpha =$). Option and item characteristic curves were re-plotted for the remaining three items and a good fit with the GRM model was demonstrated. Item information curves were plotted for the three items (Figure 44). Item #57 (Feel empty inside) demonstrated the strongest psychometric properties with respect both information and homogeneity. Overall, this revised scale appears to possess adequate scalability, discrimination, and reliability.

Negative Relationships (BOR-N)

Option and characteristic curves were plotted using Gaussian kernel smoothing techniques (Figure 45). Items #99, 139, 179, and 239 demonstrated significant violations of the assumption of monotonicity and poor fit with the GRM.

The poor psychometric function of this scale was reaffirmed by the results of the Mokken scale analysis. The scale demonstrated extremely poor fit with the Mokken model and poor reliability (scale $H = 0.13$, Rho = .46). Based upon these results the Search function of the MSP5 software was employed to determine the presence of a unidimensional central facet. A central facet was demonstrated that consisted of four items (#19, 59, 99, and 179) which demonstrated a weak but acceptable fit with the Mokken model. (scale $H = 0.37$).

Item parameters were calibrated for the 4 items found to form a central cluster (Table 22). All items displayed minimally sufficient parameter values to be maintained for further analysis, however, item #99 (people let me down) appears to best discriminate amongst respondents. Although no significant DTF was demonstrated, items #99 and 179 did demonstrate significant DIF.

Based upon item discrimination and difficulty parameters and scalability factors items #59, 99, and 179 were selected for the final revised subscale. Item parameters and scalability coefficients were re-calibrated and displayed in Table 23. The final scale demonstrated a moderate fit with the Mokken model (scale $H = 0.43$) and weak to moderate reliability (Rho = 0.68, $\alpha = 75$). Item information curves (Figure 46) indicated that item (#99) offered the most information, however all items discriminated best in the lower half of the range of theta. Given the overall findings, the final BOR-N-R must be considered a weak scale despite its superiority to the original BOR-N scale.

Self-Harm (BOR-S)

Option and item characteristic curves were plotted using Gaussian Kernel Smoothing techniques. The plots for items #263 and 343 display mild violations of the assumption of monotonicity, while items #143 and 183 did not seem to conform completely to the assumptions of the GRM (Figure 47).

Initial examination of this subscale indicated weak scalability and reliability (scale $H = .30$, Rho = .67). Further examination indicated a weak central homogeneous facet (scale $H = .37$, Rho = 0.69) consisting of items #143, 183, 223, 263, and 304. This cluster of items will be further examined through the calibration of item parameters using the GRM.

Discrimination and difficulty parameters were calibrated using the GRM (Table 23). Four of the five items demonstrated acceptable parameters, however item #263 (spend money too easy) displayed a discriminate parameter below one and an excessively low b_1 parameter. The four items with acceptable parameters were examined for the presence of DIF or DTF, however none was revealed.

The final item selection was based upon both the GRM parameters and the scalability coefficients which were largely in agreement. Item #183 (upset hurt self) offered the lowest level of information and the lowest level of scalability of the remaining items. The final scale demonstrated acceptable item parameters, scalability coefficient (scale $H = .45$), and weak reliability (Rho = .69, $\alpha = .67$). Item information curves demonstrate that item# 223 (too impulsive) offers the most information of the three items (Figure 48).

Depression Subscales

Affective (DEP-A)

Initial OCC and ICC plots computed using Gaussian kernel averaging techniques (Testgraf) for the original DEP-A subscale of the PAI as administered to a sample of MMT patients are displayed in Figure 49. Item #86 and 126 display mild violations of monotonicity, while items #166 and 206 do not conform to the assumptions regarding the hierarchical order of the response options of the GRM.

Based upon the responses of the MMT sample, the DEP-A scale was evaluated with MSP5 to determine the extent of monotone homogeneity. A weak fit was found with a scalability coefficient of 0.31 and a reliability coefficient of 0.75. Based upon these results, the scale was assessed with the Search function of MSP5 to determine if a central scale exists. Two facets were identified, the first consisted of Items #246 and #286, both of which ask the respondent to endorse the presence of happiness (scale $H = .56$, Rho = .72). A second cluster of items consisting of items #6, #46, #86, #126, #166, and #206 were found to represent a unidimensional scale (scale $H = .43$, Rho = .80). Individual

item scalability coefficients are displayed in Table 24. Given the content of the two scales, the analysis continued with the latter cluster.

The graded response model was calibrated to the remaining 6 items of the DEP-A scale. Item parameters are displayed in Table 24.The remaining 6 items were examined for DIF and DTF. A significant level of DIF was observed in item #86 (Table 24), however significant levels of DTF were not indicated.

Items #246 and #286 were eliminated as they appeared to represent a separate dimension (Happiness), while item #86 was eliminated due to the presence of DIF. Item #166 was eliminated as it offers the least amount of information of the remaining five items. The final four items were examined for fit with the Mokken model and demonstrated a moderate fit with the Mokken Model and acceptable reliability (scale *H* =. 45, Rho = .75, α = .73). Individual item parameters and calibration coefficients are displayed in Table 25. Item information curves for the four items of the DEP-A-R subscale can be viewed in Figure 50. Items # 46 (forgotten how it feels to be happy) and 126 (no pleasure) appear to offer the most information and seem to be tapping the construct of anhedonia. Total scale information curves comparing the original 8 item DEP-A with the revised 4 item version can be view in Figure 51 and demonstrates the high level of information offered by just these four items.

Cognitive (DEP-C)

OCC and ICC plots of the items of the original PAI DEP-C subscale can be viewed in Figure 52. Visual inspection of the OCC and ICC plots indicate minor violations of the assumption of monotonicity in items #227, 267, and 307. Minor

violations of the assumption of a hierarchical order of response categories were observed in most items.

The results of the plot examination were verified by the Mokken scale analysis procedure. A poor fit with the model was demonstrated and items #227, #267, and #307 exhibited item *H* parameters below the cut off of 0.30 (Table 26). The Search function was employed to determine the presence of a central facet of unidimensional items. A central facet was found to be present consisting of items #27. #67, #107, #147, and #187 which formed a scale of moderate fit (scale $H = .49$, Rho = .80). These items were maintained for the next step of analysis.

The six remaining items of the subscale were examined with Multilog to calculate the item parameters, which are displayed in Table 27. All items parameters were found to be within acceptable limits and all items were therefore examined for DIF and DTF. This analysis did not indicate the presence of either significant DIF or DTF.

The final item eliminated was done so based upon information. For that reason item #147 was eliminated leaving items #27, #67, #107, and #187 as the remaining items of the DEP-C-R subscale. Item parameters and scalability coefficients were calculated (Table 27). This revised scale demonstrated acceptable reliability ($\alpha = .73$).

The NIRT re-plotting of OCC and ICC curves indicated strong conformation to the assumption of monotonicity and improved conformation to the assumption of ordered category responses. Figure 53 displays the item information curves for the DEP-C-R subscale. Item #67, which queries feelings of worthlessness, demonstrated the most information of the remaining items. Comparison of the information offered by the new 4

item subscale as compared to the original 8 items subscale indicates no appreciable loss of information (Figure 54).

Physiological (DEP-P)

OCC and ICC plots were configure using a nonparametric kernel smoothing technique (Figure 55). Visual inspection of the OCC plots indicated minor violations of the assumption of ordered category responses of the GRM. Items #35, 115, 195, and 275 displayed minor violations of the assumption of monotonicity.

Initial analysis of the scale indicated a poor fit with the Mokken model (scale H = .25, Rho = .69). The Search function of MSP5 was employed to determine the presence of a central facet. A scale consisting of items #35, #155, #195, and #275 displayed a moderate fit with the Mokken Model (scale H = .41, Rho = .72). This scale appears to represent problems sleeping and a lack of physical energy. A second scale was found made up of two items indicating no sleep problems, however it was decided to retain the first scale based upon a superior scalability coefficient, reliability, and face validity.

The item parameters of the four remaining items were calibrated employing Multilog (Table 28). As can be seen, all items displayed acceptable item parameters. The remaining items were examined for DIF and DTF in comparison to a large sample of students. No significant DIF or DTF was demonstrated.

The final items selected for the DEP-P-R subscale consisted of the four items shown to form a Mokken scale (#35, #155, #195, and #275). This scale demonstrated a moderate fit with the Mokken model (scale H = .41, Rho = .72, α = .70). Item information curves are demonstrated in Figure 56. Item #275 (often wake up in the night) offers the highest level of information of the remaining items. Figure 57 compares the

total information offered by the original eight item scale and the new four item DEP-P-R and demonstrates the loss of minimal information given a 50% reduction in item number.

Dominance Scale (DOM)

Option and item characteristic curves were plotted using Testgraf. Visual inspection of these curves indicated that most items display violations of monotonicity (Figure 58).

The initial examination of the full DOM scale with MSP5 confirmed the lack of monotone homogeneity of the scale (scale H = .19, Rho = .71). Given this poor performance, the set of items was then examined with respect to dimensionality and three clusters were found. The first scale consisted of five items whose content centred upon seeing oneself as a leader. A second factor emerged that contained two items focusing upon a willingness to let others lead and following instruction. It was decided that the first scale dealing with perceiving oneself as a leader would be carried forward and the other items dropped.

The five remaining items of the DOM scale were examined with Multilog for the purpose of estimating item parameters which were found to be within acceptable ranges. As seen in Table 29, no significant DIF was demonstrated.

The original 12 item DOM scale was reduced to a four item scale that would seem to measure an individual's perception of themselves as a leader. The final scale displayed a moderate fit with the Mokken Model (scale H = .48, Rho=0.75, α = 76). Item scalability coefficients can be seen in Table 29. Item information curves are displayed in Figure 59. Comparison of the information provided by the original 12 item scale is compared to that of the new four item scale in Figure 60.

Drug Problems Scale (DRG)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the DRG scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 61). Visual inspection of the graphs indicated that most items demonstrate some deviations from the assumptions of monotonicity and the GRM.

The items of the PAI DRG scale were examined for fit with the Mokken model. It was found that the scale did not conform to the Mokken model (scale H = .21, Rho = .71). The Search function of the MSP5 software was employed to examine the dimensionality of the scale. A central cluster of items #22, 62, 102, 222, and 262 displayed a weak but acceptable fit with the Mokken model (scale H = .35, Rho = .69).

The graded response model was calibrated to determine item parameters. The item parameters can be seen in Table 30 and upon inspection were found to be acceptable. An analysis of differential item and test functioning of the remaining 5 items was conducting employing DFIT software based upon item parameters obtained from Multilog and Equate software, comparing the MMT population with a large sample of post-secondary school students. Results did not indicate significant levels of DIF or DTF.

The five remaining items that demonstrated conformation to the Mokken model, displayed acceptable discrimination and item parameters, and did not display significant DIF or DTF were retained as the DRG-R subscale. This revised scale demonstrated acceptable reliability (α = .68). Item information curves are displayed in Figure 62 indicating that item #62 (People tell me I have a drug problem) offers the most

information of the remaining items. As seen in Figure 63, the elimination of 7 of the original 12 items resulted in a minimal loss of information.

Mania Subscales (MAN)

Activity Level (MAN-A)

Visual inspection of the option and item characteristic curves of the MAN-A scale as plotted by Testgraf revealed significant violations of the assumptions of monotonicity (Figure 64).

Mokken scale analysis employing MSP5 supported the findings from the visual inspection of the plots (scale $H = .16$, Rho = .56). This necessitated the employment of the Search function of MSP5 to determine the dimensionality of the scale. Two scales were revealed the first consisting of items #47, #127, and #207 (scale $H = .32$, Rho = .59) and the second consisting of items #167 and #247 (scale $H = .32$, Rho = .47). The item content of the first scale seemed to be querying the presence of high levels of physical and mental activity. The second scale appeared to be measuring increased social activities and decreased need for sleep. Although both of these scales seemed to represent different symptoms and behaviours associated with manic episodes it was decided to retain the first scale based upon the psychometric features.

The item parameter coefficients were estimated for the remaining three items of the subscale as displayed in Table 31. All parameters were found to be within acceptable ranges. The remaining items were examined for DIF and DTF by comparing them to a large sample of post-secondary students. No significant DIF was demonstrated.

The MAN-A-R subscale consisted of a central facet demonstrated during the Mokken Scale Analysis which displayed a weak scalability coefficient (scale $H = .32$) and poor

reliability (Rho=0.59, α = .58). Item parameters and item scalability coefficients are displayed in Table 31. Figure 65 demonstrates the item information curves which indicate that item #207 (need to keep active not rest) offers the most information. Figure 66 compares the estimates of the total information offered by the original eight item scale and the revised four item scale indicating a loss of information. However, both the original and revised version of this scale demonstrated unacceptable characteristics.

Grandiosity (MAN-G-R)

Option and item characteristic curves were plotted for the MAN-G subscale of the PAI using nonparametric Gaussian kernel smoothing techniques (Figure 67). The results indicated that several items were not functioning in a manner consistent with the graded response model (items #68, 108, 148, 188, 228, & 308). Items #148 and 308 displayed minor violations of the assumption of monotonicity.

Initial analysis of the MAN-G subscale employed in a sample of methadone patients indicate a poor fit with the Mokken Model (scale H = .22, Rho = .67). Given these results the scale was then examined for dimensionality with the Search function of MSP5. One scale was found within the set of items consisting of items #28, 68, 108, and 188 (Table 32) which displayed barely acceptable scalability and reliability coefficients (Scale H = .34, Rho = .64).

The item parameters for the remaining four items were calibrated and found to be acceptable (Table 32). The remaining items were examined for DIF and DTF. Significant DIF was evident in items #28, 68, and 188, however, no significant DTF was demonstrated (Table 32

The final selection of items was made based upon maximizing information which resulted in the retention of items #28, #68, and #108. Item content reflected having brilliant ideas, special talents, and plans that will make the individual famous. The items were recalibrated with respect to item parameters and scalability coefficients as demonstrated in Table 32. The final three item scale demonstrated a weak fit with the Mokken Model and poor reliability (scale H = .38, Rho = .63, α = 60). Item information curves for these items are displayed in Figure 68. Figure 69 displays total subscale information curves for the original 8 item scale and the new three item scale indicating a substantive loss of information. Regardless, both versions of this subscale lack adequate psychometric properties.

Irritability (MAN-I)

Visual inspection of the OCC and ICC plots for the MAN-I subscale of the PAI employed in a sample of MMT patients did not demonstrate significant violations of the assumption of monotonicity (Figure 70). Examination of the MAN-I subscale with MSP5 indicated a weak fit with the Mokken Model (scale H = .38, Rho = .81). Individual item parameters are displayed in Table 33.

As displayed in Table 33, the item parameters of the MAN-I subscale were calibrated employing Multilog. All items demonstrated acceptable item parameters to be included in the next phase of the analysis. The eight items of the MAN-I subscale were examined for DIF and DTF employing the DFIT Framework method. As can be seen in Table 33, item #76 displayed significant levels of DIF, however no significant DTF was found at the subscale level.

Following the removal of item #76 due to DIF, the remaining items to be eliminated were done so based upon information provided. On this basis, items #116, #196, #236, and #316 were maintained as a scale which displayed a moderate scalability coefficient (scale $H = .44$) and acceptable reliability (Rho = .74, $\alpha = 71$). Re-calibrated item parameters and scalability coefficients are displayed in Table 34. Figure 71 displays the item information curves for the MAN-I-R subscale which indicates that item #236 (no patience when held back) offers the most information. Figure 72 compares total subscale information for the original eight item scale to the four item subscale which again indicates substantial loss of information. The MAN-I-R is the only subscale of the three MAN subscales that displayed acceptable.

Negative Impression Management (NIM)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the NIM scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 73). All items of this scale appear to conform to the assumption of monotonicity.

An initial analysis of the NIM scale with respect to monotone homogeneity revealed that the scale as a whole demonstrated a mild fit to the Mokken model (scale $H = .36$, Rho= 0.78). Item #249 did not conform to the assumption of monotone homogeneity (scale $H = .22$) and was eliminated from further analysis.

The graded response model was calibrated using Multilog and a summary of the discrimination and item parameter thresholds are presented in Table 35. An analysis of differential item and test functioning was conducting employing DFIT software based upon item parameters obtained from Multilog and Equate software, comparing the MMT

population with a large sample of post-secondary school students. Results indicated significant levels of DIF on item #169 (Table 35) however no significant DTF was indicated.

The final items of the NIM-R scale, following the disqualification of items #169 and #249, were chosen based upon maximizing information within this population. Therefore items #9, #89, #129, #209, and #289 were retained. The re-calibrated item parameters and scalability coefficients are displayed in Table 35. The revised scale demonstrated a weak scalability coefficient (scale $H = .38$) and acceptable reliability (Rho= 0.72, $\alpha = .71$). Item information curves for the five retained items are displayed in Figure 74, indicating that items #9 (sometimes can remember self) and #289 (no happy childhood memories) offer the most information. Figure 75 compares the total information offered by the revised version of the scale as compared to the original.

Nonsupport (NON)

Visual inspection of the option and item characteristic curves of the NON scale of the PAI as plotted using Testgraf indicated that items #121 and 281 violated the assumption of monotonicity (Figure 76).

This was supported by the results of the Mokken Scaling analysis which indicated poor scalability and reliability (scale $H = .21$, Rho = .67). The item scalability coefficients for items #121 (item $H = .14$) and 281 (item $H = .16$) were also found to be unacceptable. Therefore the scale was then examined using the Search procedure of MSP5. This second analysis indicated the presence of two central facets. The first facet consisted of items #1, 81, 161, and 201 demonstrated a moderate fit with the Mokken Model and weak reliability (scale $H = .41$, Rho = .72). This facet consisted of four

negatively worded items that appear consistent with a lack of support from family and friends. The second facet consisted of three items (#41, 241, and 281) two of which are positively worded (that appear to represent a lack of support) and one negatively worded item that appears to ask the respondent to endorse a dislike of his or her family. This scale demonstrated a barely acceptable scalability coefficient (scale H = .32) and poor reliability (Rho = .56). Given the item content and the psychometric function, the first facet was chosen to be examined with the parametric GSM. Individual item scalability coefficients are displayed in Table 37.

Item discrimination and difficulty parameters were calibrated and found to be within acceptable ranges (Table 37). Therefore it was decided that these four items demonstrated a homogeneous scale with acceptable item parameters and reliability. This revised scale demonstrated weak reliability (α = .60).

An examination of the scale employing the DFIT framework indicated no significant DIF or DTF. Item information curves were plotted and displayed in Figure 77, indicating that item #81 offered the most information.

Positive Impression Scale (PIM)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the PIM scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 78). Most items of this scale displayed minor violations of the assumption of monotonicity, however Item #344 seriously violated this assumption. This item asked the respondent to deny the presence of good moods.

Initial analysis of this scale indicated a poor fit with the Mokken model (scale *H* = .23, Rho=0.72). In order to assess the dimensionality of the scale the Search mode of MSP5 was employed. The analysis indicated one major scale comprised of all the item of the scale except for item #344. These items were maintained in the continuing analysis.

Following the removal of Item #344, the graded response model was calibrated using Multilog and a summary of the discrimination and item parameter thresholds are presented in Table 36. An analysis of differential item and test functioning was conducting employing DFIT software based upon item parameters obtained from Multilog and Equate software, comparing the MMT population with a large sample of post-secondary school students. Results indicated significant levels of DIF on items #104 and #264 (Table 36); however there was no significant DTF.

Table 38 displays the items retained from the original PIM scale for use in the assessment of MMT populations and their item parameters. This revised scale demonstrated acceptable reliability (α = .71). Information curves were computed for each item of the final set of the most effective items of the PIM scale of the PAI based upon responses from the MMT population (Figure 79). As can be viewed in Figure 80, the newly revised PIM-R scale for use in MMT population offers maximum information at approximately θ = -0.5 and the total test information provided by the new 5-item PIM-R scale is compared to the initial estimates of total test information for the original 9-item PIM scale.

Paranoia (PAR)

Hypervigilance (PAR-H)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the PAR-H scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 81). In the case of this subscale the majority of the items displayed minor violations to the assumption of monotonicity. However, the majority of these violations tended to occur within the extreme ranges of theta.

Examination of the PAR-H scale using MSP5 indicated a very poor fit with the model (scale H = .11, Rho=0.48). Based upon these results, the items were examined for dimensionality with the Search function of MSP5. Two scales were indicated each of which contained two items. The first scale, which consisted of items #88 and #128, displayed a moderate fit with the Mokken model (scale H = .47, Rho=0.64). This scale appears to be measuring the extent to which an individual trusts others. A second weaker scale displayed a mild fit with the Mokken model (scale H = .36, Rho=0.49) and consisted of items #48 and #168. The item content appears to be surveying thoughts of betrayal and hidden motives. This two item scale appears to be more consistent with the purpose of the original scale and the concept of hypervigilance; therefore item parameters were calculated for this potential two item scale.

The graded response model was calibrated for the two remaining items (Table 39) with acceptable item parameters obtained. This revised scale demonstrated weak reliability (α = .48).An analysis of differential item and test functioning was conducted employing DFIT software based upon item parameters obtained from Multilog and Equate software, comparing the MMT population with a large sample of post-secondary school students. Significant DIF was demonstrated; however significant DTF

was not (Table 39). Overall, this scale demonstrated unacceptable psychometric properties. Figure 82 displays item information curves for these two items.

Persecution (PAR-P)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the PAR-P scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 83). In the case of this subscale most items displayed minor violations of the assumptions of IRT models, however, item #148 and item #229 displayed serious violations of the assumptions of monotonicity.

The PAR-P subscale was examined employing the Test function of the MSP5 and as indicated in the initial NIRT examination the scale exhibited a poor fit with the Mokken Model (scale H = .22, Rho=0.66). Further examination using the Search function of the MSP5 software indicated the presence of one central scale consisting of items that exhibited a moderate fit with the Mokken Model (scale H = .42, Rho=0.77).

The graded response model was calibrated to the remaining 5 items of the PAR-H subscale using Multilog software. A summary of the discrimination and difficulty parameters are presented in Table 40. An analysis of differential item and test functioning was conducting employing DFIT software based upon item parameters obtained from Multilog and Equate software, comparing the MMT population with a large sample of post-secondary school students. Results did not indicate significant DIF (Table 40) or DTF with any of the remaining items.

Given the acceptable psychometrics of all the retained items, it was decided to retain the items offering the maximum amount of information when used within the

methadone population. Therefore items # 69, #149, #189, and #269 were retained, while item #29 was eliminated. Item parameters and scalability coefficients were re-calibrated for the scale and can be viewed in Table 40. The scale displayed a moderate fit with the Mokken Model (scale H = .44, Rho=0.75, α = .74). Information curves were computed for each item of the final set of the most effective items based upon responses from the MMT population (Figure 84). As can be viewed in Figure 85 the removal of 4 items from the original 8 resulted in minimal information loss.

Resentment (PAR-R)

Option and item characteristic curves of the PAR-R scale were plotted using Testgraf based upon the responses from the MMT sample Figure 86. Visual inspection indicated that most of the items violated the assumption of monotonicity.

Examination of the PAR-R subscale using the Mokken modeling software MSP5 indicated a poor fit (scale H = .13, Rho= 0.54). These results indicated the need to examine the dimensionality of the subscale. Such an examination was conducted and resulted in the suggestion of three different subscales. The first factor, which displayed the strongest fit with the Mokken Model (scale H = .53, Rho= 0.69) contained two items which when endorsed appear to represent an individual's perception that his or her efforts are not being appreciated by others. A second weaker scale (scale H = .35, Rho=0.58) that consisted of three reverse scored items that seem to reflect the perception of being treated fairly by others. A third similarly weak scale consisting of two items seemed to reflect the perception that success and respect are earned. Given the original intent of the scale and the general lack of homogeneity, it was decided to maintain the first scale reflecting resentment.

The two items found to represent a scale measuring resentment were calibrated to the GRM as demonstrated in Table 41. The two item scale was examined to determine if significant DIF or DTF was present. Significant DTF was not exhibited nor was significant DIF (Table 41). This revised scale demonstrated acceptable reliability (α = .49).The two item scale was maintained as PAR-R-R as part of the revised PAI for use in methadone maintenance treatment populations. Figure 87 displays item information curves for the two items of the PAR-R-R subscale.

Treatment Rejection (RXR)

Option and item characteristic curves were plotted using Testgraf (Figure 88). Visual inspection of the plots that items #42 displayed a minor violation of monotonicity while item #282 displayed significant violations.

Initial analysis of this scale for fit with the Mokken model indicated a weak fit (scale H = .25, Rho = .71). Therefore the Search function of the MSP5 software was employed to determine if a dominant facet exists within this set of items. This analysis indicated the presence of a four item central factor which displayed a mild fit with the Mokken Model (scale H = .43, Rho=0.72). The scale consisted of items #2. #42, #82, and #122. Item content for this group seems to centre upon the need for change and the need for help to change. A second weaker scale (scale H = .35, Rho=0.52) was demonstrated consisting of items #202 and #282. The content of these items seem to represent the belief that one can solve one's own problems and that one is comfortable with one's self. For that reason the first four item scale will be maintained given the slightly stronger psychometric properties and the better fit with the intended purpose of the original scale.

The remaining four items of this scale were examined with Multilog in order to calibrate item discriminatory and difficulty parameters (Table 42). All items demonstrated acceptable parameters and were carried forward in the analysis. Significant DIF was demonstrated in items #2, #42, and #122 (Table 42), however, no significant DTF was demonstrated.

Since all analyses indicated acceptable properties for the four item scale indicated by the Mokken Scale analysis it was decided that the revised scale should include these items. This revised scale demonstrated weak reliability ($\alpha = .69$). Item information curves were plotted and are displayed in Figure 89. As can be seen the item which offers the most information (Item #122) asks the respondent to endorse the need for help. Figure 90 compares the total information offered by the original eight item scale as compared to the revised version. This graph demonstrates that these four items offer the majority of the information as offered by the original.

Schizophrenia (SCZ)

Psychotic Experience (SCZ-P)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANT S scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 91). Visual inspection of the OCC and ICC plots indicated a strong likelihood that many of the items of the SCZ-P scale violate the assumptions of monotonicity and homogeneity.

Initial examination of the SCZ-P scale indicated a poor fit and a lack of monotone homogeneity (scale $H = .21$, Rho=0.61). Examination of the scale with the Search function of MSP5 indicate a central scale consisting of items #10, #50, #170, and #210

that demonstrated a weak but acceptable fit with Mokken model (scale H = .35, Rho= 0.65). Item content indicated the endorsement of strange thoughts and hearing voices. It was decided that this four item scale would be further analyzed.

The discrimination and difficulty parameters were estimated for the remaining items using Multilog (Table 43). All items demonstrated discrimination parameters above the minimum level of one and were retained for analysis within the DFIT framework. An analysis of differential item and test functioning was conducting employing DFIT software employing item parameters obtained from Multilog and Equate, comparing the MMT population with a large sample of post-secondary school students. Item #10 from the original SCZ P scale was shown to display significant DIF. There was no significant DTF demonstrated.

Table 43 displays the items retained for the SCZ-P-R from the original SCZ P scale for use in the assessment of MMT populations. Information curves were computed for each item of the SCZ-P-R scale of the PAI based upon responses from the MMT population and are display in Figure 92. Figure 93 displays the Total Information Curves for the new SCZ-P-R subscale (4 items) as compared to the original SCZ P subscale (8 items). This revised scale demonstrated weak reliability (α = .64).

Social Detachment (SCZ-S)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the SCZ S scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 94). All items on this scale displayed violations of the assumption of monotonicity.

The items of this scale were examined using MSP5. The scale displayed a poor fit with the model (scale H = .21, Rho=0.66). An examination of the scale employing the Search function of MSP5, to determine if a central facet exists was conducted. Two main central facets were demonstrated each consisting of four items. The first cluster consisted of items #30 and 70 displaying a moderate scalability coefficient (scale H = .38) and acceptable reliability (Rho=0.51). Item content of this scale reflected being a loner and not feeling close to others. A second cluster was found (Items #190, #230, #270, and #310) that also displayed a moderate fit with the Mokken Model (scale H = .47) with acceptable reliability (Rho=0.76). These items reflect a sociability factor. Based upon item content it was decided to retain the first factor since its item content better matches the intended purpose of the subscale.

. Item parameters for the two remaining items were calibrated using the Multilog software and are displayed in Table 44. No significant DIF or DTF was demonstrated in the remaining two items. Item information curves for the remaining two items are displayed in Figure 95. This revised scale demonstrated weak reliability (α = .46).

Thought Disorder (SCZ-T)

Option characteristic curves (OCC) and item characteristic curves (ICC) were computed for each item of the ANT S scale of the PAI, based upon the responses of the methadone maintenance treatment (MTT) population (Figure 96). Visual inspection of the plots indicates that most items conform to the assumption of monotonicity (with some minor irregularities), however Item #318 does not.

Initial analysis of the scale employing the MSP5 software indicated a weak fit with the model (scale H = .34, Rho=0.77). All items demonstrated acceptable scalability coefficients with the exception of Item #318. The Search function was employed and confirmed the earlier findings indicating that the seven items of the original scale other than item #318, form a cluster of moderate fit with the Mokken Model (scale H = .43, Rho=0.82). Item scalability coefficients are displayed in Table 45.

The discrimination and difficulty parameters were estimated for the remaining items using Multilog (Table 45). An analysis of differential item and test functioning was conducting employing DFIT software employing item parameters obtained from Multilog and Equate, comparing the MMT population with a large sample of post-secondary school students. This analysis revealed that Item #198 displayed significant NCDIF; however significant levels of Differential Test Functioning were not detected (Table 45).

Table 46 displays the items retained for the SCZ-T-R from the original SCZ T scale for use in the assessment of MMT populations. Information curves were computed for each item of the SCZ-T-R scale of the PAI based upon responses from the MMT population and are display in Figure 97. Figure 98 displays the Total Information Curves for the new SCZ-T-R subscale (4 items) as compared to the original SCZ T subscale (8 items). This revised scale demonstrated acceptable reliability (α = .78).

Somatic (SOM)

Conversion (SOM-C)

The SOM-C scale of the PAI was initially examined using the nonparametric item response theory based software Testgraf (Figure 99). Visual inspection of the plots revealed no violations of the assumptions of the GRM.

The eight items of the SOM-C subscale were examined for monotone homogeneity using MSP5 software. This examination revealed a mild fit with the model (scale H = .39, Rho=0.81) with all items meeting the minimum scalability coefficient requirements.

Following an initial examination of the SOM C subscale, all items of the scale were examined with Multilog. It was found that Item #3 lacked discriminatory powers, and was removed from the list of potential items for this scale. A second such examination revealed that all items had acceptable discrimination and difficulty parameters to be included in the next stage of the analysis (Table 47). The remaining seven items of the SOM C scale were analyzed with DFIT to estimate the extent of DIF or DTF, following the calculation of linking parameters. No significant DIF or DTF was identified (Table 47).

The SOM C scale performed very well during all phases of this analysis. For that reason, decisions regarding item retention were based upon the item parameters in order to maximize the information provided by the subscale. The retained items and re-calibrated item parameters and scalability coefficients are displayed in Table 48. A Mokken scale analysis indicated a moderate fit (scale H = 0.44, Rho=0.73, α = 0.74). Individual item information curves are displayed in Figure 100, while Figure 101 compares total information curves from the original SOM-C subscale and the revised version.

Health Concerns (SOM-H)

The original items of the SOM-H subscale were examined for monotonicity through a nonparametric IRT analysis using Testgraf. The ICCs and OCCs for each item

are displayed in Figure 102. Visual inspection of the plots indicated that item # 292 violates the assumption of monotonicity. Minor irregularities were displayed by items #12, #92, #172, #212, and #252.

Analysis of this scale with respect to monotone homogeneity indicated a weak fit with the Mokken model (scale H = .34, Rho = .77), while at the item level #292 was found to have an item level scalability coefficient of 0.15. Given this additional evidence of the ineffectiveness of Item #292 it was eliminated from further consideration.

An initial estimation of item parameters was conducted with Multilog examining the 7 remaining items of the SOM-H subscale. These parameters indicated that Item #172 offered insufficient information, characterized by a low discrimination parameter (below 1) and a wide range of difficulty parameters. This item was dropped from the subscale and item parameters were re-estimated and all found to be within acceptable ranges (Table 49). The item parameters and scores of each group on the remaining six items were linked using EQUATENP. The obtained linking parameters were employed to determine whether significant DIF or DTF was present. As seen in Table 49, a significant level of DIF was demonstrated in Item #252, however no significant DTF was observed.

Given the lack of significant item bias, the final four item of this scale were chosen based upon the quality of information offered. Based upon the item parameters obtained, it was decided that Items #12 and #252 (also displayed significant DIF) offered the least information and were eliminated. The scale now consisted of four items (#52, #92, #132, and #212). A final examination of these four items was conducted to arrive at final item parameters (Table 50) and estimates of item and test information (Figure 103 and Figure 104). This revised scale demonstrated weak reliability (α = .68).

Somatization (SOM-S)

The original items of the SOM-S subscale were examined for monotonicity through a nonparametric IRT analysis using Testgraf. The ICCs and OCCs for each item are displayed in Figure 105. Results indicated that Item #152 should be eliminated from the subscale prior to analysis with Multilog.

The SOM-S subscale was examined for monotone homogeneity. Two separate scales were revealed, the first of which consisted of four items (#32, #72, #112, and #192) displaying a mild fit with the model (scale H = .39, Rho=0.70). A second scale was evident that consisted of two items and also displayed a weak fit with the Mokken Model (scale H = .35, Rho=0.51). It was decided that the four item scale would be retained for further analysis.

As can be seen in Table 51, all four remaining items demonstrated acceptable parameters. The remaining six items were then examined for the presence of DIF and DTF. Items #112 and #272 were removed due to significant levels of DIF however no significant DTF was not demonstrated (Table 51). This revised scale demonstrated acceptable reliability (α = .80). This revised scale demonstrated weak reliability (α = .64). Given the result of this analysis, the remaining four items were retained for the new SOM-S-R. Item information curves for each item are displayed in Figure 106. Figure 107 compares the estimated information offered by the original SOM-S subscale and the revised version.

Stress Scale (STR)

The option characteristic curves and item characteristic curves for the original eight items of the STR scale of the PAI were plotted using Testgraf based upon responses

from a sample of methadone maintenance treatment patients (Figure 108). Item #326 displayed significant violations of the assumption of monotonicity.

The STR scale was examined for monotone homogeneity employing the MSP5 software Test feature. A poor fit with the model was found (scale H = .25, Rho = .70). The scale was then examined using the Search function to identify scales that conform to the assumptions of monotone homogeneity. One central facet was identified consisting of seven items (scale H = .41, Rho = .80). This analysis also indicated the removal of item #326.

The item parameters for the remaining seven items of the STR scale were calibrated using Multilog. As can be seen in Table 52, all items displayed acceptable parameters. The remaining seven items of the STR were examined for DIF and DTF. As seen in Table 52, significant DIF was demonstrated in six of the seven remaining items. Significant DTF was indicated (DTF=0.486, chi square=586.74, p=.000). Following the removal of item #324 no significant DTF remained.

Six items remained that displayed acceptable scalability coefficients, item parameters, and did not contribute to significant levels of DTF. The final four items were chosen based upon maximizing information. This set of items consisted of items #321, #322, #323, and #325. These four remaining items were re-examined for monotone homogeneity, item parameters were recalibrated, and item and scale information curves were plotted (Table 53, Figure 109, and Figure 110). This revised scale demonstrated acceptability reliability (α = .74).

Suicidal Ideation (SUI)

The results of the NPIRT analysis of the SUI scale in the MMT population indicated that all items displayed monotonicity, with the exception of Item #341 (Figure 111). As well, visual inspection of the OCC indicated that the item generally conformed to the graded response model.

A Mokken Scale analysis was performed on the 12 items of the original SUI scale of the PAI. A moderate fit with the Mokken scale was found (scale H = .48, Rho=0.89). However, the scalability coefficient for Item #341 was found to be lower than the recommended cut off value of 0.30. This item was eliminated from further consideration.

Following the removal of Item #341 from the SUI scale, the discrimination and difficulty parameters were estimated for the remaining items using Multilog (Table 54). It was found that the difficulty parameter for Item #301 was below one. This item was removed and the parameters were re-estimated. In this analysis, all remaining item discrimination parameters were greater than one. The item parameters from the remaining ten items were used to calculate linking parameter using EQUATENP. With these linking parameters, the DFIT framework was used to examine these remaining items for DIF and DTF. No significant DIF or DTF was indicated.

Following the elimination of items due to a lack of monotonicity, a lack of information, or the presence of significant DIF or DIF, ten items of the original twelve items of the SUI scale of the PAI remained. The final six items were chosen based upon maximizing the information provided by the scale. The retained items tended to focus upon current and past cognitions regarding suicide (Table 55). Figure 112 depicts the individual item information curves for the items retained in the 6-item SUI-R scale. Figure 113 compares the total test information for the 6-item SUI-R scale as compared to

the original 12-item SUI scale. This revised scale demonstrated strong reliability (α = .90).

Warmth Scale (WRM)

Visual inspection of the items of the WRM subscale indicated that many items were not consistent with the assumptions of the IRT models. Items #13, 93, 133, 173, 213, 293, 330, and 331 displayed minor violations of monotonicity, while item # 332 displayed more serious difficulties (Figure 114). Items #133, 173, 213, 253, 330, and 332 did not conform to the hierarchical option assumption of the GRM. Many items on this scale displayed a non-zero lower asymptote indicating that even individuals who scored at the lowest level of the underlying trait were most likely to choose an option other than 0.

The original WRM subscale was analyzed using the MSP5 software and three separate scales were demonstrated. The strongest scale consisted of two items and appeared to be representing the inability to quickly develop warm friendships (scale H = .54, Rho=0.70. The second scale of moderate fit which ask the respondent to endorse being sociable and warm (scale H = .41, Rho=0.71). A third weak scale was indicated that seemed to be querying the extent to which a respondent is sympathetic and affectionate towards others (scale H = .38, Rho=0.69). It was decided that the four item scale querying a sociability warmth factor would be maintained for the next step of analysis.

All four remaining items of the WRM scale were found to exhibit acceptable item discrimination and difficulty parameters (Table 56). Given these results, these remaining

items were examined for the presence of DIF and DTF. This analysis did not reveal the presence of DIF (Table 56) nor was DTF detected.

Based upon the results of this analysis, these four items were retained for inclusion in the WRM-R scale. Item information curves are displayed in Figure 115, while Figure 116 compares the total information offered by the original scale as compared to the revised version. The most information offered by this scale was provided by Item #93 which asked respondents to endorse the attribute of enjoying meeting new people. OCC and ICC were re-plotted with Testgraf and these plots indicated that all four items now demonstrated a good fit with the GRM, good discrimination across the continuum, and monotonicity. This revised scale demonstrated acceptable reliability (α = .71).

Discussion

This extensive IRT examination of the PAI based upon the responses of individuals being treated in two separate methadone maintenance programs resulted in the derivation of a potential revised version of the PAI for use in substance abusing populations. Alternatively, these results offer an improved scoring algorithm for the original PAI when used in such populations. This potential inventory is comprised of 155 items that were chosen based upon a systematic examination employing both nonparametric and parametric item response theory models and methods. Items were retained based upon the formation of homogeneous scales or subscales, maximizing the information offered by each scale or subscale (a combination of discriminatory strength and range of difficulty parameters) and minimizing DIF and or DTF. In general, the item content was not used as a method of determining item retention except when choosing between facets of a

scale demonstrated to be multidimensional. However, if this algorithm were employed to create a completely new test based upon original untested items, validity would be an important factor. In this case, such work was completed previously by the developer of the PAI. The majority of the new scales developed in this research offer similar levels of information as the originals despite a 50% reduction in the number of items. In other words, discriminatory power was largely retained, but much error was discarded. This reduction in item number should substantially reduce completion time. Therefore, it might also be hypothesized (but not possible to examine within the context of this research) that the actual administration of this potential version of the PAI to an adequate sample of MMT patients might demonstrate further psychometric benefits. In such clinical populations there are many threats to test validity such as difficulties with low frustration tolerance, attention and focus that a shorter test may help reduce.

Many scales, subscales, and items of the PAI demonstrated strong psychometric function as examined through IRT methods. The scales and subscales examining symptoms of depression, symptoms of anxiety, and symptoms of alcohol and substance abuse demonstrated strong psychometric function as did the suicidal ideation and the trauma related scales. However, it was found that several PAI scales and subscales consist of multiple factors that adversely affect the interpretability of the total scale or subscale score. Particular psychometric weaknesses were demonstrated in the Anxiety Related Disorder subscales, the Mania subscales, the Schizophrenia subscales, and the Paranoia subscales. These poor results largely mirror the results of Boyle and Lennon (1994) discussed earlier, despite the disparate methodologies between these two studies. With respect to bipolar disorder, anxiety disorders (beyond PTSD), psychotic disorders,

or personality disorders (other than Antisocial and Borderline), it seems imperative that clinicians employing the PAI as a screening tool or as an aid in diagnosis closely examine the individual responses to items on elevated scales to better determine their meaning.

Multidimensionality

This analysis revealed several potential sources of multidimensionality. For example, the anxiety related disorders scales of compulsive behaviours and phobias are not specific in nature. For instance, the ARD-P subscale mixes symptoms from several Axis I anxiety related disorders. The effective items of this scale sample content related to social phobia. The other items of this scale tend to examine specific phobias that do not form a homogeneous scale. Therefore, in the PAI-R, this scale is reduced to the only two items that form a unidimensional scale assessing social phobia.

Another potential source of multidimensionality is the attempt to simultaneously measure symptoms of Axis I and Axis II disorders. With the exception of Antisocial and Borderline Personality Disorders, the items that measure symptoms and behaviours associated with personality disorders are imbedded within scales of subscales that also measure Axis I disorders. This can result in multidimensional scales and scores that are difficult to interpret. For example, the ARD-O subscale queries symptoms and behaviours from the Axis I obsessive compulsive disorder (OCD) and Axis II symptoms of obsessive compulsive personality disorder (OCPD). Recent research has indicated that OCD and OCPD do not exist on the same continuum. Crino and Andrews (1996) conducted a study that examined personality disorders in a sample of individuals being treated for OCD and other anxiety disorders in a tertiary care anxiety clinic in Australia. There findings indicated that individual's diagnosed with OCD are no more likely to be

diagnosed with an Axis II disorder than individuals diagnosed with other anxiety disorders. More salient was the finding that while traits consistent with OCPD are found in those diagnosed with OCD; similar levels of such traits are found in other Axis I anxiety disorders. This empirical evidence supports the contention that symptoms associated with Axis II disorders should not be included on scales intended to be used to aid in the diagnosis of Axis I disorders.

Multidimensional scales and subscales can also be the result of response bias and method effects. Acquiescence response bias is the tendency for respondents to give similar or identical responses to a series of items that use the same response format and direction of wording. It has been suggested that the use of both negatively and positively (straight forwardly) worded items will reduce such bias. In such cases negatively worded items are reverse scored in order to place them on the same continuum as the positively worded items. Unfortunately the employment of negatively worded items introduces additional threats to the psychometric function of a scale such as multidimensionality due to method effects.

Method effects might be defined as artifactual effects of a specific test construction method. In the case of negatively worded items, analyses typically employing factor analytic methods have consistently demonstrated that such items load on a factor separate from positively worded items. Most authors suggest that this apparent multidimensionality is in reality an artifactual method effect introduced through the use of negatively worded items. Several reasons for this effect have been suggested ranging from such items being difficult to understand or a tendency to respond to such items in a different manner. A third possible explanation for such effects is the introduction of a

true separate dimension. For example, Rodebaugh, Woods, and Heimburg (2007) concluded that the negatively worded reverse-scored items of the Social Interaction Anxiety Scale are actually measuring extraversion, which although correlated with social anxiety, is a separate dimension. Their extensive analysis led these researchers to conclude that the reverse scored items hindered the psychometric function of the scale. As a result Rodebaugh et al. recommended that scoring of the SIAS exclude the reverse scored items. Similar investigations have been conducted with the Rosenberg Self Esteem Scale (Gana, Alaphilippe, & Bailly, 2005; Greenberger, Chen, Dmitrieva, & Farruggia, 2003) and the Brief Fear of Negative Evaluation Scale – Revised (Carleton, McReary, Norton, & Asmundson, 2006; Carleton, Collimore, & Asmundson, 2007). In most cases the results indicate that negatively worded items result in method effects that impair the psychometric function of the scale with the tendency to underestimate the severity of the intended trait.

In a study combining both CTT and IRT methodologies to examine the Fear of Negative Evaluation scale (FNE; Watson & Friend, 1969) and the Brief Fear of Negative Evaluation scale (BFNE; Leary, 1983), it was found that not only did reverse worded items represent a separate factor, but these items consistently offered less information as well (Rodebaugh, Woods, Thissen, Heimburg, Chambless, & Rapee, 2004). Reverse scored items typically ask the respondent to endorse behaviours and cognitions hypothesized to be on the same continuum as psychopathological construct of interest but at the opposite pole. Rodebaugh et al. hypothesize that such items likely confuse the respondent rather than represent a separate construct; however, they do indicate that their analysis was unable to determine the probability of correctness of either hypothesis. The

methodology differed from that employed in this dissertation. Dimensionality was examined through the use of factor analysis which was then followed by the calibration of item parameters based upon the factors. These factors were found to divide the scales by the type of wording (straightforward or reverse). NIRT methods were not employed at all in this analysis. The entire scale was not examined as a whole via IRT methods. It appears that the authors decided a priori to only examine the scales with IRT methods stratified by wording (straight forward versus reverse).

Approximately one third of the items of the PAI are negatively worded and reverse scored. The literature search for this dissertation did not reveal the examination of large inventories such as the PAI with respect to such method effects, nor was this subject an original area of interest in this research. However, the results of this IRT analysis of the PAI certainly bring this issue front and centre. The resulting version of the PAI contains only one negatively worded item. In most cases such items were the first removed from a scale due to the demonstration of poor psychometric function. In some cases the negatively worded items were shown to form a unique facet, in which case true multidimensionality must be given serious consideration. These results suggest that the widespread use of negatively worded items in the PAI impairs the psychometric function of its scales and potentially limits the estimation of the constructs of interest. It further suggests that if such items are truly necessary to reduce acquiescence bias, that they not be used in the scoring of the scales as recommended by Rodebaugh et al.

Information

A major concern in the reduction of the number of items contributing to a scale would be a potential reduction in information. This would reduce the precision of a scale and its

ability to discriminate between respondents. This concern however is largely based upon the assumption that each item is effective and is contributing to the total information. However, in the case of ineffective items the major contribution may in fact be error. As well, if a scale contains items that are ineffective, such items often do not conform to the assumptions of the model being employed to calculate information levels and may actually result in the overestimation of the information offered by the whole scale. If ineffective items offering little information are removed from a scale, significant levels of error are also being removed from the scale. Once such items are removed truer estimates of the information offered by a scale can be estimated.

DIF and DTF

The PAI proved to be particularly sound with respect to differential item functioning and more importantly differential test functioning. This IRT analysis of the PAI was noteworthy given the general lack of significant findings of DTF. Based upon the recommendations of IRT theorists discussed earlier, these findings certainly suggest that most items of the PAI are well constructed in this respect and support the utility of the PAI in a variety of settings (Hidalgo-Montesinos and Lopez-Pina, 2002; Collins, Raju, and Edwards, 2000). The entire original PAI as administered to the two groups of interest was initially examined for DIF and DTF prior to the elimination of items that violated the assumptions of the GRM. At this exploratory stage many items were found to display significant DIF, however very few scales were found to display DTF. However, when the algorithm for choosing items was employed, the majority of items that did display DIF were eliminated prior to examination for DIF and DTF. In fact only one item was eliminated due to significant DTF in the final analysis that resulted in the revised PAI. It

seems that in the identification of ineffective items that lack discrimination or information from the results of either IRT model also serves to eliminate DIF and DTF. Given the large computational and time costs of the DFIT analysis, these results raise questions regarding the utility of the conducting such analyses, in that this portion of the analysis so rarely paid dividends. On the other hand this could speak to the quality of the scales and items of the PAI. Like many other procedures, the use of the DFIT framework does not provide yes or no decisions regarding which items to include or exclude. It requires subjective decision making that should be completed in the context of the overall IRT examination, including other factors such as information, discrimination, and conformation to the assumptions of the IRT model in question.

Efficiency and Effectiveness of IRT Algorithms

Not only did this research provide a rich source of information regarding the PAI and its employment in substance abusing populations, it also demonstrated the utility of employing both parametric and nonparametric item response theory models and analyses. A high level of agreement was found between all methods employed. Correlations between the GRM discrimination parameters for the final set of items and the item scalability factors (item *H*) was found to significant ($R = .83$, $p < .001$). This significant correlation is especially meaningful, given that the sample size in this research (while adequate for NIRT analyses) was minimally sufficient for PIRT.

These results provide additional support to the importance of integrating both IRT and CTT methods in the construction of measures of psychopathology including personality disorders. For instance the PAI is a widely employed measure of such constructs that demonstrates adequate psychometric properties from a CTT perspective.

However, from an IRT perspective some areas of concern were demonstrated such as multidimensional scales and items offering low information that can best be identified through the employment of both NIRT and PIRT methods. Given the widespread availability of IRT software (despite their previously discussed user unfriendliness), future construction of specific and broad measures of psychopathology should employ IRT methods in conjunction with CTT methods. Furthermore, IRT methods offer an untapped avenue of investigation into method effects associated with negatively worded and positively worded items in a unique manner.

Initially each scale and subscale was analyzed by each specific NIRT and PIRT method in isolation and again afterwards in the sequential manner reported in the results section. Regardless of the order in which these analyses took place, the results generally were the same. All methods tended to identify the same items as ineffective. In short, if such an analysis had been conducted only using Mokken Scaling Analysis or employing only the Graded Response Model (via Multilog) the resulting revised scales would be extremely similar. Although, it has been recommended that a thorough IRT analysis of should begin with an NIRT screening for monotonicity etc., it appears that eliminating items demonstrating insufficient information can also deal effectively with this issue.

It must be stated that the number of analyses, numbers of softwares employed, and the computational cost of this dissertation was prohibitive. In terms of maximizing both efficiency and effectiveness in the analysis of polytomous data examining psychopathology and personality, the results of this dissertation would seem to support the combination of an initial examination of assumptions with MSP5 followed by an item calibration employing PIRT software such as Multilog as suggested by many IRT

researchers (Ramsay, 2000; Meijer and Baneke, 2004; Stout, 2001) This would likely result in a monotone homogeneous scale that maximizes information for the population in question and the efficiency of the process. Although Multilog does not specifically identify items which violate the assumption of monotonicity, these analyses revealed that such items generally have discrimination parameters well below one and wide ranging difficulty parameters and are as ineffective. As well, the results of these analyses would indicate that this process would likely eliminate items that display DIF contributing to DTF.

With respect to the employment of Testgraf, the plotting of ICCs and OCCs offers both benefits and drawbacks. These plots allow visual inspection to determine if the items (and options) of a scale are conforming to the hierarchical assumptions of the GRM and to examine for monotonicity without forcing the data to fit a specified function before and after revision. After the derivation of the revised scales and subscales, OCC and ICC were re-plotted for each revised scale or subscale based only upon the retained items. In all cases, the new plots demonstrated superior monotonicity, discrimination, and adherence to the hierarchical option assumptions of the GRM. This result indicates that not only did the purification process conducted result in improved shorter scales, but NIRT regression was able to detect such improvements. However, visual inspection of items based upon the graphical output of Testgraf proved difficult in items with mild violations of monotonicity when employed with the full scales. As well, Testgraf does not generally provide quantitative measures of item functioning to aid in the retention or elimination of items.

Rescaling and Rescoring

Original NIRT plotting of OCC and ICC of the scales and subscales of the PAI with Testgraf tended to indicate significant rescoring of the PAI would be in order, however, following the suggested revisions and re-plotting of these curves it was found that revised scales demonstrated a much better fit with the hierarchical option assumptions of the GRM. Test developers and psychometricians who suggest the rescoring of items as a solution to poor item function should keep in mind that option characteristic curves, item characteristic curves, item parameters, and model fit are not only dependent upon the item itself but also upon the scale to which they are included. For instance, if a scale is multidimensional, many OCC and ICC plots will indicate that an item is quite ineffective as a member of such a scale. However, if the dimensionality of the scale is examined and the items are separated into proper clusters, the model fit and plots may improve tremendously. For example Figure 117 demonstrates the option and item characteristic curves of several items of the ARD-O subscale before and after rescaling. These results indicate that often rescaling will eliminate the need to rescore items. However in some cases, for example the ALC-R scale, even following scale purification, the items still do not conform to the GRM (Figure 118).

Future Research

The next step in the validation of this abbreviated version of the PAI would most certainly be the comparison of the construct validity of the new scales and subscales as compared to the old. This could be accomplished through the examination of data that includes the concurrent administration of the PAI and other well validated measures of psychopathology. In this manner, the correlation rates of the old and new subscales (for example the ARD-O and the ARD-OCD) with a short scale such as the Obsessive

Compulsive Inventory (Foa, Huppert, Leiberg, Langner, Kichic, Hajcak, & Salkovskis, 2002). It is suggested that employing such high quality scales as criterion measures would be superior to employing scales from large inventories such as the MMPI-2 as done in the original validation studies of the PAI.

The creation of improved software packages for the both NIRT and PIRT analysis of likert style self-report measures of personality and psychopathology is not only needed but essential for the continued improvement of such measures. The most obvious need is for the improvement of software for the conduction of differential item and test functioning. At this point the most promising method of such analysis, the DFIT framework, requires the employment of three separate software applications and the creation and copying of several types of data files. Realistically, the process to complete such analyses is beyond the grasp of many researchers. The analysis for this dissertation involved hundreds of hours of computer time, the creation of hundreds of data files, the creation of hundreds of output files, and the employment of six different IRT software packages each of which required a significant amount of time to learn. The development of a software program that would incorporate all or some of these capabilities within one software application would be of obvious benefit and importance.

Although the Testgraf software has limitations given its lack of quantitative output, it does contain a feature that if incorporated into other types of analysis would prove highly informative. Testgraf give the research the ability to examine the function of a scale items using another score as the estimate of the underlying trait as opposed to the raw scale score. An example of such an analysis using a substance abusing population will be given to help illustrate this concept. In most methadone clinics, patients are

required to submit regular urine samples for toxicology analysis. Such an analysis can determine not only the presence of substances of abuse, but in some cases can determine the extent of use. Such analyses can easily be converted to a total score based on the number of substances of abuse, the extent of use, and the seriousness of substance. For instance, heroin would be considered more serious than cannabis and would receive a higher score on that parameter. The function of the items of the DRG scale would then be examined based upon the score received from the urine toxicology results. In this manner, the items from the PAI DRG scale which best form a scale to predict actual urine toxicology results within a specific population could be determined.

Computerized adaptive testing, based upon IRT methods, has been shown to be extremely efficient within academic testing environments such as in the administration of standardized tests like the Graduate Record Exam (GRE), the Law School Aptitude Test (LSAT), and the SAT. However, little use has been made of such applications within the field of measurement of psychopathology despite empirical evidence of its utility (Waller and Reise, 1989; Gardner, Shear, Kelleher, Pajer, Mammen, Buysse, and Frank, 2004). However, the results of this dissertation and the demonstrated stability of item parameters would seem to indicate the potential of such methods with respect to the PAI. Future efforts should examine the feasibility of such applications in the measurement of psychopathology.

Given the unexpected finding of multidimensionality on many scales and subscales of the PAI, an extensive examination of the PAI employing a multidimensional IRT model may be of benefit. For example, if the endorsement of an item relies upon more than one latent trait, the use of such a model allows for the concurrent estimation of

multiple traits (Embretson & Reise, 2000). Such models may demonstrate a better fit with the existing scales of the PAI.

Limitations of this Research

Based upon the literature reviewed for this dissertation, the number of subjects in the clinical sample was minimally sufficient (Lautenschlager, Meade, & Kim, 2006; Reise and Yu, 1990) for use with PIRT methods. This could have resulted in unstable parameter estimates during the parametric analysis. A second limitation of this sample size issue was the inability to directly assess local independence.

This study does not directly address the issue of construct validity. While the scales of the PAI have most likely been improved psychometrically and in many cases item clusters to be retained were chosen based upon face validity, it is necessary to rely upon the previously conducted validity studies during the development of the PAI regarding such validity issues. Future research should include studies that examine the concurrent validity of the revised PAI clinical scales and subscales developed in this dissertation. This could likely be accomplished through the re-examination of the existing data from previous validity studies and recalculating the scale scores based upon the items suggested by this dissertation. As well, new investigations might for example compare the concurrent administration of both the ARD-OCD-R and the ARD-OCPD-R and a quality OCD scale such as the Obsessive Compulsive Inventory – Revised and the SCID-II scale querying Obsessive Compulsive Personality Disorder (Foa et al., 2002).

Further IRT examination of the PAI would benefit from the examination of large samples of individuals diagnosed with specific psychiatric disorders. For example a similar algorithm as employed in this dissertation might be repeated with a large clinical

sample (mood and anxiety disorders, psychotic disorders) where the level of specific underlying traits would be greater. This would allow for further examination of some scales that did not perform well in this analysis (e.g., MAN, SCZ). It is possible that some items and scales may perform more effectively (i.e., offer more information) if employed in samples likely to demonstrate higher levels of the underlying trait beyond substance use. Using similar logic, it would have been beneficial to have been able to examine differential item and test functioning in a larger clinical sample in which other well established psychiatric diagnoses were known.

Summary

The PAI is a widely employed measure of psychopathology and abnormal personality validated for use in a variety of populations as a screening tool, diagnostically, and in treatment planning. It has been examined and found psychometrically sound through the employment a variety of CTT methods. Through the employment of a variety of IRT methods a shorter version of the PAI was derived which offers similar levels of information as the original with over 50% less items. The results of this dissertation did demonstrate some weaknesses with the original PAI from an IRT perspective. The most salient issue would be the multidimensionality of many of the scales, followed by the presence of many ineffective items that offer little in the way of information to discriminate between respondents. The item selection procedure employed for this dissertation used both NIRT and PIRT methods to identify items to retain and those to discard. The results highlight the utility of IRT methods in the construction of measures of psychopathology and abnormal personality and by definition the examination of how each item functions within the sample.

The PAI is an example of an inventory that was constructed using traditional methods and assessed psychometrically with CTT methods (Morey, 1991). The results of this dissertation demonstrate the benefits of including IRT analyses in the construction of polytomously scored inventories examining psychopathology. This is especially relevant given Morey's philosophy of creating items that focus upon capturing the respondent's conceptualization of the relevant behaviour or symptom as it relates to the objective diagnostic criteria. Such a philosophy and resultant method of item construction is a worthy objective; however, it necessitates the thorough examination of these items beyond relying upon total scale scores and their correlations with other criterion measures.

References

Alterman, A., Cacciola, J., Habing, B., and Lynch, K. (2007). ASI recent and lifetime summary indices based on nonparametric IRT models. *Psychological Assessment, 19,* 119-132.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorder,* Fourth Edition. Washington, DC : Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorder,* Fourth Edition, Text Revision. Washington, DC : Author.

Baker, Frank (2001). *The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation,* University of Maryland, College Park, MD.

Boyle, G.J. and 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.

Bradley, R., Hilsenroth, M., Guarnaccia, C., and Westen, D. (2007). Relationship between clinician assessment and self-assessment of personality disorders using the SWAP-200 and PAI. *Psychological Assessment, 19* (2), 225-229.

Cacciola, J.S., Alterman, A.I., Rutherford, M.J., McKay, J.R., and Mulvaney, F.D. (2001). The relationship of psychiatric comorbidity to treatment outcomes in methadone maintained patients. *Drug and Alcohol Dependence, 61,* 271-280.

Calhoun, P.S., Sampson, W.S., Bosworth, H.B., Feldman, M.E., Kirby, A.C., Hertzberg, M.A., Wampler, T.P., Tate-William, F., Moore, S.D., and Beckham, J.C. (2000). Drug use and validity of substance use self-reports in veterans seeking help for

posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology, 68,* 923-927.

Carleton, R.N., McCreary, D.R., Norton, P.J., and Asmundson, G.J.G. (2006). Brief Fear of Negative Evaluation Scale-Revised. *Depression and Anxiety, 23,* 297-303.

Collins, W.C., Raju, N.S., and Edwards, J.E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451-461.

Crino, R.D. and Andrews, G. (1996). Personality disorders in obsessive compulsive disorder: A controlled study. *Journal of Psychiatric Research, 30,* 1, 29-38.

De Koning, E., Sijtsma, K., and Hamers, J.H.M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement, 26,* 3, 302-320.

Edens, J.F., Hart, S.D., Johnson, D.W., Johnson, J.K., and Olver, M.E. (2000). Use of the Personality Assessment Inventory to assess psychopathy in offender populations. *Psychological Assessment, 12* (2), 132-139.

Edens, J.F., Poythress, N.G., and Watkins-Clay, M.M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: a comparison of the PAI, SIMS, and SIRS. *Journal of Personality Assessment, 88* (1), 33-42.

Eisen, S.V., Normand, S., Belanger, A.J., Spiro, A., and Esch, D. (2004). The BASIS-R: Reliability and Validity. *Medical Care, 42*, 1230-1241.

Embretson, S. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341-349.

Embretson, S. and Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ : Lawrence Erlbaum Associates.

Flowers, C.P., Oshima, T.C., and Raju, N.S. (1999). A description and demonstration of Polytomous-DFIT framework. *Applied Psychological Measurement, 23* (4), 309-326.

Foa, E.B., Huppert, J.D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., and Salkovskis, P.M. (2002). The Obsessive-Compulsive Inventory: development and validation of a short version. *Psychological Assessment, 14* (4), 485-496.

Gardner, W., Shear, K., Kelleher, K.J., Pajer, K.A., Mammen, O., Buysse, D., and Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *Psychiatry, 4* (13), 1-11.

Gana, K., Alaphilippe, D., and Bailly, N. (2005). Factorial structure of the French version of the Rosenberg Self-Esteem Scale among the elderly. International Journal of Testing, 5 (2), 169-178.

Greenberger, E., Chen, C., Dmitrieva, J., and Farruggia, S.P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences, 35,* 1242-1254.

Guy, L.S., Pythress, N.G., Douglass, K.S., and Skeem, J.L. (2008). Correspondence between self-report and interview-based assessments of antisocial personality disorder. Psychological Assessment, 20 (1), 47-54.

Haller, D.L., Miles, D.R., and Dawson, D.S. (2002). Psychopathology influences treatment retention among drug-dependent women. *Journal of Substance Abuse Treatment, 23,* 431-436.

Hare, R.D. (1991). *The Hare Psychopathy Checklist-Revised.* Toronto: Multi-Health Systems.

Hart, S.D., Cox, D.N., and Hare, R.D. (1995). *Manual for the Screening Version of the Psychopathy Checklist (PCL:SV).* Toronto: Multi-Health Systems.

Hidalgo-Montesinos, M.D. and Lopez-Pina, J.P.(2002). Two-stage equating in differential item functioning detection under the graded response model with Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62,* 32-44.

Hogan, D.M. (1998). Annotation: the psychological development and welfare of children of opioid and cocaine users: review and research needs. *Journal of Child Psychology and Psychiatry, 39,* 609-620.

Iwata, N., Turner, R.J., and Lloyd, D.A. (2002). Race/ethnicity and depressive symptoms in community-dwelling young adults: a differential item functioning analysis. *Psychiatry Research, 110,* 281-289.

Jacobo, M.C., Blais, M.A., Baity, M.R., and Harley, R. (2007). Concurrent validity of the Personality Assessment Inventory borderline scales in patients seeking dialectical behavior therapy. *Journal of Personality Assessment, 88* (1), 74-80.

Junker, B.W. and Sijtsma, K. (2001) Nonparametric Item Response Theory in Action: An Overview of the Special Issue. *Applied Psychological Measurement, 25,* 211–220.

Karlin, B.E., Creech, S.K., Grimes, J.S., Clark, T.S., Meagher, M.W., and Morey, L.C. (2005). The Personality Assessment Inventory with chronic pain patients:

psychometric properties and clinical utility. *Journal of Clinical Psychology, 61* (12), 1571-1585.

Kaye, S., Darke, S., and Finlay-Jones, R. (1998). The onset of heroin use and criminal behaviour: does order make a difference? *Drug and Alcohol Dependence, 53*, 79-86.

Kellog, S.H., Ho, A., Bell, K. Schluger, R.P., McHugh, P.F., McClary, K.A., and Kreek, M.J. (2002). The Personality Assessment Inventory drug problems scale: A validity analysis. *Journal of Personality Assessment, 79* (1), 73-84.

Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U., Kendler, K.S., 1994. Lifetime and 12-months prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry, 51* (1), 8–19.

Krause, M., Degkwitz, P., Kuhne, A., and Verthein, U. (1998). Comorbidity of opiate dependence and mental disorders. *Addictive Behaviors, 23*, 767-783.

Lautenschlager, G. J., Meade, A. W., and Kim, S.-H. (2006, April). *Cautions Regarding Sample Characteristics When Using the Graded Response Model.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Leary, M.R. (1983). A brief version of the Fear of Negative Evaluation scale. *Personality and Social Psychology Bulletin, 9,* 371-375.

Ledgerwood, D.M. and Downey, K.K. (2002). Relationship between problem gambling and substance use in a methadone population. *Addictive Behaviors, 27*, 483-491.

Lesieur, H.R. and Blume, S.B. (1987).The South Oaks Gambling Screen (SOGS): a new instrument for the identification of pathological gamblers. *American Journal of Psychiatry, 144*, 1184-1188.

McDevitt-Murphy, M.E., Weathers, F.W., Adkins, J.W., and Daniels, J.B. (2005). Use of the Personality Assessment Inventory in assessment of posttraumatic stress disorder in women. *Journal of Psychopathology and Behavioral Assessments, 27* (2), 57-65.

McDevitt-Murphy, M.E., Weathers, F.W., Flood, A.M., Eakin, D.E., and Benson, T.A. (2007). The utility of the PAI and MMPI-2 for discriminating PTSD, depression, and social phobia in trauma-exposed college students. *Assessment, 14* (2), 181-195.

McDonald, R.J. (1999). *Test theory: a unified treatment.* Mahwah, NJ : Lawrence Erlbaum Associates.

Meijer, R.R and Baneke, J.J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Mokken, R.J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York: De Gruyter.

Molenaar, I.W. (1991). A weighted Loevinger H coefficient extending Mokken scaling to multicategory items. *Kwantitative Methoden, 12*, 97-117.

Molenaar, I.W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25,* 295-299.

Morales, L.S., Flowers, C., Gutierrez, P., Kleinman, M., and Teresi, J.A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care, 44*, 143-151.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional Manual.* Odessa, FL: Psychological Assessment Resources.

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

Morey, L.C. & Lanier, V.W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment, 5* (3), 203-214.

Mueser, K.T., Noordsy, D.L., Drake, R.E., and Fox, L. (2003). *Integrated treatment for dual disorders: a guide to effective practice.* New York: Guilford Press.

Nowaka, J.A., Roberson-Naya, R., Strong, D.R., Bucceria, J., and Lejeuz, C.W. (2003). Using item response theory in the development and validation of the College-Oriented Eating Disorders Screen. *Eating Behaviors, 4*, 345–361.

Ouimette, P.C., Brown, P.J., and Najavits, L.M. (1998). Course and Treatment of patients with both substance use and posttraumatic stress disorder. *Addictive Behaviours, 23,* 785-795.

Parker, J.D., Daleiden, E.L., and Simpson, C.A. (1999). Personality Assessment Inventory substance use scales: convergent and discriminatory relations with the Addiction Severity Index in a residential chemical dependence treatment setting. *Psychological Assessment, 11* (4), 507-513.

Raju, N.S., van der Linden, W.J., and Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19* (4), 353-368.

Ramsay, J.O. (2000). *Testgraf, A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data,* McGill University.

Regier, D.A., Farmer, M.E., Rae, D.S., Locke, B.Z., Keith, S.J., Judd, L.L., and Goodwin, F.K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchments Area (ECA) Study. *Journal of the American Medical Association, 264* (19), 2511-2518.

Reise, S.P, Ainsworth, A.T., and Haviland, M.G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*, 95-101.

Reise, S.P. and Yu, J. (1990). Parameter recovery in the Graded Response Model using Multilog. *Journal of Educational Measurement, 27*, 133-144.

Risser, D., Honigschnable, S., Stichenwirth, M., Pfudle, S., Sebald, D., Kaff, A., and Bauer, G. (2001). Mortality of opiate users in Vienna, Austria. *Drug and Alcohol Dependence, 64,* 251-256.

Rodebaugh, T.L., Woods, C.M., and Heimburg, R.G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do no belong. *Behavior Therapy, 38,* 192-206.

Rodebaugh, T.L., Woods, C.M., Thissen, D.M., Heimburg, R.G., Chambless, D.L., and Rapee, R.M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16* (2), 169-181.

Ruiz, M.A., Dickinson, K.A., and Pincus, A.L. (2002). Concurrent validity of the Personality Assessment Inventory Alcohol Problems (ALC) scale in a college student sample. *Assessment, 9* (3), 261-270.

Samejima, F. (1967). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

Santor, D. and Ramsay, J.O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10*, 345-359.

Santor, D.A. and Coyne, J.C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9,* 3, 233-243.

Santor, D.A., Ramsay, J.O., and Zuroff, D.C. (1994). Nonparametric item analyses of the Beck Depression Inventory evaluating gender item bias and response option weights. *Psychological Assessment, 6*, 255-270.

Schinka, J.A. (1995). Personality Assessment Inventory scale characteristics and factor structure in the assessment of alcohol dependency. *Journal of Personality Assessment, 62* (1), 101-111.

Sijtsma, K., Debets, P., and Molenaar, I.W. (1990). Mokken scale analysis for polychotomous items: theory, a computer program and an empirical application. *Quality & Quantity, 24*, 173-188.

Stein, M.B., Pinsker-Aspen, J.H., and Hilsenroth, M.J. (2007). Borderline pathology and the Personality Assessment Inventory (PAI): an evaluation of criterion and concurrent validity. *Journal of Personality Assessment, 88* (1), 81-89.

Steinburg, L. and Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*_(1), 81-97.

Tasca, A., Wood, J., Demidenko, N., and 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*, 337–356.

Thissen, D. and Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Torrey, W.C., Drake, R.E., Cohen, M., Fox, L.B., Lynde, D., Gorman, P., and Wyzik, P. (2002). The challenge of implementing and sustaining integrated dual disorders treatment programs. *Community Mental Health Journal, 38*, 607-521.

Uttaro, T. and Lehman, A. (1999). Graded response modeling of the Quality of Life Interview. *Evaluation and Program Planning, 22,* 41-52.

Van Schuur, W.H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*, 139-163.

Waller, N.G. and Reise, S.P. (1989). Computerized adaptive personality assessment: An illustration with the *Absorption Scale. Journal of Personality and Social Psychology, 57,* 1051-1058.

Wang, E.W., Rogers, R., Giles, C.L., Diamond, P.M., Herrington-Wang, L.E., and Taylor, E. R. (1997). A pilot study of the Personality Assessment Inventory (PAI) in corrections: assessment of malingering, suicide risk, and aggression in male inmates. *Behavioral Sciences and the Law, 15,* 469-482.

Watson, D., and Friend, R. (1969). Measurement of social evaluative anxiety. Journal of *Consulting and Clinical Psychology, 33,* 448-457.

Appendix A

Revised Version of the PAI for Use in MMT Populations

SCALE/SUBSCALE	PAI#	CONTENT
AGG-A	258	Bad temper
AGG-A	338	Hard to calm down
AGG-A	339	People think I am aggressive
AGG-P-R	21	People afraid of my temper.
AGG-P-R	61	Temper explodes, lose control.
AGG-P-R	101	Sometimes very violent.
AGG-V-R	58	Tell people off when deserved
AGG-V-R	138	Yell to get point across
ALC-R	15	Sometimes feel guilty
ALC-R	55	Trouble controlling use
ALC-R	95	Have had to cut down
ALC-R	135	Causes relationship problems
ALC-R	215	People think I drink too much
ALC-R	255	Causes problems at home
ANT-A-R	51	deliberately damaged property
ANT-A-R	91	done things that weren't legal
ANT-A-R	131	used to lie a lot
ANT-A-R	171	see how much can get away with
ANT-E R	71	Take advantage of others
ANT-E R	111	Do most things for a price
ANT-S-R	39	Kick out of dangerous things
ANT-S-R	79	Do wild things for fun
ANT-S-R	119	Behaviour is wild.
ANT-S-R	279	Doesn't turn down a dare.
ANX-A-R	4	Extremely tense
ANX-A-R	44	Nervousness
ANX-A-R	84	Afraid for no reason
ANX-A-R	204	Feel something terrible
ANX-C-R	25	Trouble concentrating, nervous
ANX-C-R	65	Hard to enjoy because of worry
ANX-C-R	105	So worried can't stand it
ANX-C-R	265	Worry more than necessary

ANX-P-R	73	Worry so much, feels like fainting
ANX-P-R	113	Dizzy when under pressure
ANX-P-R	153	Often feel heart pounding
ANX-P-R	233	Trouble breathing under pressure
ARD-O-R	5	Do things a certain way
ARD-O-R	45	Fight to control impulses
ARD-O-R	85	Bothered by things out of place
ARD-P-R	26	Fear of saying something wrong
ARD-P-R	66	Exaggerated fears
ARD-P-R	106	Nervous in front of others
ARD-T-R	34	Reliving horrible event
ARD-T-R	114	Troubled by bad memories
ARD-T-R	154	Can't get over something
ARD-T-R	234	Nightmares about past
BOR-A-R	14	Sudden mood shifts
BOR-A-R	54	Intense mood shifts
BOR-A-R	214	Can't do enough to express anger
BOR-I-R	17	Attitude about self changes
BOR-I-R	57	Feel empty inside
BOR-I-R	137	Wonder what to do with life
BOR-N-R	59	Let people know they hurt me.
BOR-N-R	99	People let me down.
BOR-N-R	179	Picked wrong friends
BOR-S-R	143	Impulsivity causes trouble
BOR-S-R	223	Too impulsive
BOR-S-R	303	Reckless
DEP-A-R	6	Often sad no reason
DEP-A-R	46	Forgotten feeling happy
DEP-A-R	126	No pleasure
DEP-A-R	206	No interest in life
DEP-C-R	27	Let everyone down
DEP-C-R	67	Worthless
DEP-C-R	107	Don't feel like trying
DEP-C-R	187	Nothing works

DEP-P-R	35	Hardly any energy
DEP-P-R	155	Moving more slowly
DEP-P-R	195	Wake up early, can't sleep
DEP-P-R	275	Often wake up in night
DOM-R	16	Take charge person
DOM-R	56	Natural leader
DOM-R	96	Good at telling others
DOM-R	176	Best as a leader
DRG-R	22	Use to feel better
DRG-R	62	Told have a drug problem
DRG-R	102	Caused financial strain
DRG-R	222	Use is out of control
DRG-R	262	Caused health problems
INF-R	40	Favourite poet
INF-R	80	Ads in mail
INF-R	120	Favourite sports event
INF-R	160	Rather win
INF-R	200	Favourite hobbies
INF-R	240	Overpriced
INF-R	280	Look forward to dentist
INF-R	320	Free time
MAN-A-R	47	Take on too many commitments
MAN-A-R	127	Racing thoughts
MAN-A-R	207	Need to keep active, not rest
MAN-G-R	28	Many brilliant ideas
MAN-G-R	68	Very special talents
MAN-G-R	108	Plans will make me famous
MAN-I-R	116	Upset when others don't understand
MAN-I-R	196	Bothered by those who don't understand
MAN-I-R	236	No patience with being held back
MAN-I-R	316	No patience for those who disagree
NIM-R	9	Cannot remember who I am.
NIM-R	89	Destined to unhappiness
NIM-R	129	Multiple personalities
NIM-R	209	Lose memory
NIM-R	289	No happy childhood memories

PIM-R	24	little things bother me
PIM-R	104	complain too much
PIM-R	144	too impatient
PIM-R	224	put things off
PIM-R	264	make promises I can't keep
PAR-H-R	48	Alert to unfaithful people
PAR-H-R	168	People hide real motives
PAR-P-R	69	People make me look bad
PAR-P-R	149	Some people try to keep me from getting ahead.
PAR-P-R	189	People keep me from getting ahead
PAR-P-R	269	People have it in for me.
PAR-S-R	117	Given a lot, not much in return
PAR-S-R	157	People don't appreciate me
RXR-R	2	Inner struggles cause problems.
RXR-R	42	Need to make changes.
RXR-R	82	Need to change even if it hurts.
RXR-R	122	Need help to deal with problems.
SCZ-P-R	10	Ideas others think strange.
SCZ-P-R	50	People put thoughts into my head.
SCZ-P-R	170	Heard voices no one else heard.
SCZ-P-R	210	People try to control my thoughts.
SCZ-S-R	30	Don't relate to people
SCZ-S-R	70	Don't have much to say
SCZ-T-R	38	confused thinking
SCZ-T-R	78	scrambled thoughts
SCZ-T-R	118	trouble keeping thoughts separate.
SCZ-T-R	278	thoughts disappear
SOM-C-R	83	Numbness unexplained
SOM-C-R	123	Double or blurred vision
SOM-C-R	163	Eyesight got worse then better
SOM-C-R	243	Legs so weak can't walk
SOM-H-R	52	Complicated health problems
SOM-H-R	92	Medical problems cause struggle
SOM-H-R	132	Difficult to treat medical issues
SOM-H-R	212	Unusual diseases and illnesses

SOM-S-R	32	Often don't feel well
SOM-S-R	72	Suffer from pain
SOM-S-R	112	Good health
SOM-S-R	192	Bad back
STR-R	321	Money problems
STR-R	322	Life is unpredictable.
STR-R	323	Many life changes
STR-R	325	Family problems
SUI-R	60	thought about ways to kill self.
SUI-R	100	made plans how to kill self
SUI-R	140	Recently thinking about suicide
SUI-R	180	Thought about suicide long time
SUI-R	220	Death would be a relief.
SUI-R	260	Thought about suicide note
WRM-R	13	Very sociable
WRM-R	53	Making friends is easy
WRM-R	93	Like to meet new people
WRM-R	133	Warm person

Table 1

Alpha coefficient calculated for the PAI and the PAI-R scales and subscales based upon responses from various clinical populations

	PAI (344 items)				PAI-R (172 items)
	Clinical Standardization	Chronic Pain ($N = 432$)	Alcohol Treatment	MMT	MMT
			Validity Scales		
Positive Impression	0.77	0.45	0.79	0.70	0.77
Negative Impression	0.74	0.75	0.60	0.78	0.71
			Clinical Scales and Subscales		
Somatic Complaints	0.92	0.85	0.90	0.89	0.87
Conversion	0.83	0.74	0.77	0.80	0.74
Somatization	0.77	0.62	0.70	0.76	0.80
Health Concerns	0.83	0.76	0.81	0.65	0.68
Anxiety	0.94	0.93	0.92	0.89	0.91
Cognitive	0.87	0.86	0.85	0.72	0.81
Affective	0.84	0.79	0.81	0.69	0.76
Physiological	0.83	0.81	0.78	0.74	0.73
Anxiety-Related Disorder	0.86	0.83	0.83	0.83	0.88
Obsessive-Compulsive	0.63	0.64	0.62	0.61	0.64
Phobias	0.67	0.62	0.56	0.46	0.70
Traumatic Stress	0.89	0.84	0.89	0.89	0.70

Table 1 (cont.)

		PAI (344 items)				PAI-R (172 items)
		Psychiatric Standardization	Chronic Pain ($N = 432$)	Alcohol Treatment	MMT	MMT
Depression		0.91	0.91	0.91	0.87	0.88
	Cognitive	0.84	0.83	0.80	0.71	0.79
	Affective	0.88	0.90	0.86	0.74	0.73
	Physiological	0.80	0.75	0.76	0.68	0.70
Mania		0.80	0.80	0.81	0.83	0.76
	Activity Level	0.55	0.53	0.59	0.56	0.58
	Grandiosity	0.78	0.73	0.75	0.67	0.60
	Irritability	0.81	0.80	0.82	0.81	0.71
Paranoia		0.84	0.84	0.89	0.74	0.72
	Hypervigilance	0.75	0.71	0.74	0.47	0.48
	Persection	0.83	0.65	0.81	0.66	0.74
	Resentment	0.72	0.67	0.70	0.50	0.49
Schizophrenia		0.89	0.83	0.89	0.80	0.79
	Psychotic Experience	0.71	0.48	0.68	0.61	0.64
	Social Detachment	0.85	0.82	0.88	0.65	0.46
	Thought Disorder	0.85	0.83	0.85	0.77	0.78

Table 1 (cont.)

	PAI (344 items)				PAI-R (172 items)
	Psychiatric Standardization	Chronic Pain ($N = 432$)	Alcohol Treatment	MMT	MMT
Borderline Features	0.91	0.87	0.90	0.86	0.87
Affective Instability	0.81	0.77	0.84	0.64	0.72
Identity Problems	0.77	0.73	0.66	0.70	0.76
Negative Relationships	0.68	0.68	0.66	0.45	0.75
Self-Harm	0.76	0.65	0.79	0.64	0.67
Antisocial Features	0.86	0.84	0.84	0.80	0.81
Antisocial Behaviours	0.80	0.65	0.71	0.60	0.60
Egocentricity	0.63	0.45	0.64	0.68	0.48
Stimulus Seeking	0.75	0.53	0.74	0.69	0.75
Alcohol Problems	0.93	0.73	0.75	0.82	0.88
Drug Problems	0.89	0.58	0.92	0.70	0.68
Treatment Consideration Scales and Subscales					
Aggression	0.90	0.88	0.90	0.82	0.85
Aggressive Attitude	0.80	0.74	0.80	0.67	0.71
Verbal Aggression	0.70	0.67	0.86	0.75	0.79
Physical Aggression	0.84	0.71	0.70	0.42	0.67
Suicidal Ideation	0.93	0.89	0.93	0.90	0.90

Table 1 (cont.)

	PAI (344 items)				PAI-R (172 items)
	Psychiatric Standardization	Chronic Pain Pain	Alcohol Treatment	MMT	MMT
Stress	0.79	0.80	0.79	0.69	0.74
Nonsupport	0.80	0.77	0.76	0.60	0.60
Treatment Rejection	0.80	0.75	0.60	0.69	0.69
		Interpersonal Scales			
Dominance	0.82	0.79	0.80	0.68	0.76
Warmth	0.83	0.79	0.83	0.64	0.71

Note: Data for the Clinical Standardization group are taken from Morey (1991), for the Chronic Pain group from Karlin et al. (2005), and for the Alcohol Dependent group from Schinka (1995). PAI = Personality Assessment Inventory. MMT = methadone maintenance treatment.

Table 2

Example of a perfect Guttmann polytomous scale with 6 items (four options each).

	Most Severe ⟷ Least Severe						
Response Pattern	Item#1	Item#2	Item#3	Item#4	Item#5	Item#6	Frequency
1	0	0	0	0	0	0	25
2	0	0	0	0	0	1	25
3	0	0	0	0	0	2	1
4	0	0	0	0	1	2	2
5	0	0	0	1	2	2	1
6	0	0	0	1	2	3	1
7	0	0	1	2	2	3	17
8	0	0	1	2	3	3	18
9	0	1	2	2	3	3	2
10	0	1	2	3	3	3	3
11	0	2	2	3	3	3	1
12	1	2	3	3	3	3	2
13	2	2	3	3	3	3	1
14	3	3	3	3	3	3	1

Table 3

Item Parameters and scalability coefficients for the PAI AGG-A subscale in the MMT population ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
			AGG-A[a]			
258	Bad temper	2.29 (0.24)	-0.37 (0.09)	0.52 (0.10)	1.12 (0.12)	0.40
298	temper never causes problems	0.89 (0.15)	-2.42 (0.43)	-1.25 (0.26)	0.38 (0.19)	0.33
299	Anger never out of control	0.87 (0.15)	-2.41 (0.44)	-1.06 (0.24)	0.49 (0.20)	0.32
338	Hard to calm down	1.93 (0.20)	-0.58 (0.11)	0.40 (0.10)	1.11 (0.14)	0.37
339	People think I am aggressive	1.33 (0.18)	-0.31 (0.13)	0.66 (0.14)	1.56 (0.22)	0.31
			AGG-A-R			
258	Bad temper	2.18 (0.23)	-0.38 (0.09)	0.52 (0.10)	1.13 (0.12)	0.48
338	Hard to calm down	2.17 (0.22)	-0.56 (0.10)	0.38 (0.09)	1.05 (0.13)	0.49
339	People think I am aggressive	1.56 (0.19)	-0.29 (0.12)	0.60 (0.12)	1.43 (0.18)	0.45

[a]Following the removal of items #298 and 299.

Table 4

Item Parameters and scalability coefficients for the items of the PAI AGG-P subscale following the removal of item #221 and the AGG-P-R subscale in the MMT population (N = 323).

Item	Content	*a* *SE*	*b1* *SE*	*b2 SE*	*b3 SE*	Item *H*
				AGG-P[a]		
21	People afraid of my temper.	1.70 (0.21)	-0.04 (0.11)	0.80 (0.13)	1.62 (0.19)	0.45
61	Temper explodes, lose control.	3.26 (0.36)	-0.19 (0.07)	0.63 (0.07)	1.05 (0.10)	0.57
101	Sometimes very violent.	3.33 (0.37)	0.20 (0.07)	1.04 (0.09)	1.41 (0.11)	0.55
141	Smash things when I'm upset.	1.56 (0.20)	0.15 (0.12)	0.99 (0.14)	1.81 (0.22)	0.44
181	Threatened to hurt people.	1.45 (0.21)	0.25 (0.12)	1.33 (0.19)	1.92 (0.26)	0.42
				AGG-P-R		
21	People afraid of my temper.	1.66 (0.20)	-0.04 (0.11)	0.81 (0.13)	1.65 (0.20)	0.54
61	Temper explodes, lose control.	4.78 (0.54)	-0.18 (0.06)	0.59 (0.06)	0.98 (0.07)	0.65
101	Sometimes very violent.	2.62 (0.30)	0.20 (0.08)	1.09 (0.11)	1.50 (0.14)	0.62

[a]Following the removal of item #221.

Table 5

Item Parameters for the two items of the PAI AGG-V-R subscale in the MMT population (N = 323).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
58	Tell people off when deserved	2.10 (0.21)	-1.46 (0.14)	-0.19 (0.09)	0.60 (0.10)	0.52
138	Yell to get point across	2.13 (0.23)	-1.20 (0.12)	-0.13 (0.08)	0.64 (0.10)	0.52

Table 6

Examination of the 12 items of the PAI ALC scale in a sample of MMT patients (N = 323).

Item #	Content	Mean	Item H
15	Sometimes feel guilty	0.46	0.52
55	Trouble controlling use	0.32	0.55
95	Have had to cut down	0.89	0.45
135	Causes relationship problems	0.50	0.55
175	Helps in social situation	0.55	0.41
215	People think I drink too much	0.44	0.50
254	Drink first thing in the morning	0.27	0.42
255	Causes problems at home	0.55	0.52
294	Never drink and drive	1.45	0.23
295	Hardly ever drink	1.34	0.36
334	Never gotten into trouble	2.02	0.35
335	Caused problems at work	0.55	0.42

Table 7

Item parameters and scalability coefficients for the ALC scale following the removal of item #294 in a MMT patient sample (N = 323).

Item#	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
15	Sometimes feel guilty	2.93 (0.44)	0.69 (0.08)	1.06 (0.11)	1.36 (0.13)	0.55
55	Trouble controlling use	3.35 (0.54)	0.85 (0.08)	1.31 (0.12)	1.62 (0.18)	0.57
95	Have had to cut down	2.54 (0.32)	0.12 (0.09)	0.59 (0.10)	0.88 (0.11)	0.50
135	Causes relationship problems	4.62 (0.65)	0.62 (0.07)	0.90 (0.07)	1.13 (0.09)	0.57
175	Helps in social situation	2.07 (0.28)	0.46 (0.10)	1.06 (0.14)	1.67 (0.21)	0.43
215	People think I drink too much	3.16 (0.40)	0.64 (0.09)	1.15 (0.11)	1.47 (0.15)	0.53
254	Drink first thing in the morning	2.01 (0.38)	1.17 (0.16)	1.77 (0.24)	1.99 (0.29)	0.44
255	Causes problems at home	3.94 (0.53)	0.56 (0.07)	0.85 (0.07)	1.12 (0.09)	0.55
295	Hardly ever drink	0.80 (0.18)	-0.35 (0.22)	0.13 (0.22)	0.86 (0.30)	0.35
334	Never gotten into trouble	0.63 (0.18)	-2.03 (0.58)	-1.45 (0.42)	-0.68 (0.28)	0.36
335	Caused problems at work	2.22 (0.32)	0.62 (0.11)	1.01 (0.13)	1.41 (0.18)	0.45

Table 8

Item parameters and scalability coefficients for the ALC scale following the removal of item #294 in a MMT patient sample (N = 323).

Item#	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item H	*NCDIF*
15	Sometimes feel guilty	2.94 (0.48)	0.62 (0.08)	0.98 (0.11)	1.27 (0.14)	0.54	0.011
55	Trouble controlling use	3.31 (0.56)	0.77 (0.08)	1.23 (0.13)	1.54 (0.18)	0.57	0.007
95	Have had to cut down	2.73 (0.36)	0.06 (0.08)	0.50 (0.09)	0.78 (0.10)	0.58	0.013
135	Causes relationship problems	4.97 (0.81)	0.54 0.07)	0.81 (0.06)	1.04 (0.09)	0.61	0.010
175	Helps in social situation	2.20 (0.31)	0.39 (0.10)	0.96 (0.14)	1.54 (0.20)	0.47	0.118
215	People think I drink too much	3.20 (0.44)	0.57 (0.08)	1.06 (0.11)	1.38 (0.15)	0.55	0.010
254	Drink first thing in the morning	2.03 (0.40)	1.09 (0.16)	1.68 (0.24)	1.90 (0.30)	0.46	0.005
255	Causes problems at home	4.07 (0.59)	0.49 (0.07)	0.77 (0.07)	1.03 (0.09)	0.58	0.024
335	Caused problems at work	2.33 (0.35)	0.54 (0.10)	0.91 (0.13)	1.30 (0.18)	0.47	0.026

Table 9

Item parameters and scalability coefficients for the ALC-R scale in a MMT patient sample ($N = 323$).

Item#	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
15	Sometimes feel guilty	2.97 (0.43)	0.77 (0.08)	1.13 (0.11)	1.44 (0.13)	0.60
55	Trouble controlling use	3.42 (0.52)	0.93 (0.08)	1.39 (0.12)	1.70 (0.17)	0.63
95	Have had to cut down	2.58 (0.32)	0.20 (0.08)	0.66 (0.09)	0.96 (0.10)	0.63
135	Causes relationship problems	4.71 (0.87)	0.71 (0.07)	0.98 (0.07)	1.22 (0.09)	0.64
215	People think I drink too much	3.10 (0.40)	0.72 (0.08)	1.22 (0.11)	1.56 (0.16)	0.58
255	Causes problems at home	3.65 (0.46)	0.65 (0.08)	0.95 (0.08)	1.22 (0.10)	0.60

Table 10

Item parameters and scalability coefficients for the ANT-A-R scale based upon responses from a MMT sample ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	Item *H*
51	deliberately damaged property	1.32	(0.22)	1.16	(0.18)	1.95	(0.29)	2.65	(0.42)	0.33
91	done things that weren't legal	1.37	(0.19)	-1.58	(0.20)	-0.21	(0.12)	0.45	(0.13)	0.41
131	used to lie a lot	1.47	(0.19)	-0.81	(0.14)	0.54	(0.12)	1.14	(0.16)	0.35
171	see how much can get away with	1.48	(0.20)	0.28	(0.11)	1.36	(0.18)	2.04	(0.26)	0.35

Table 11

Item parameters and scalability coefficients for the ANT-E subscale ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
71	Take advantage of others	1.64 (0.25)	0.65 (0.12)	1.77 (0.22)	2.22 (0.29)	.37
111	Do most things for a price	1.45 (0.18)	-0.26 (0.12)	0.74 (0.13)	1.79 (0.22)	.37

Table 12

Item parameters and scalability coefficients for the five remaining items of the ANT-S subscale after the NIRT analysis ($N = 323$) and the ANT-S-R subscale.

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
				ANT-S[a]		
39	Kick out of dangerous things	2.06 (0.26)	0.41 (0.09)	1.30 (0.14)	1.88(0.21)	0.42
79	Do wild things for fun	3.00 (0.35)	0.19 (0.07)	1.01 (0.09)	1.55 (0.13)	0.44
119	Behaviour is wild.	2.52 (0.28)	-0.07 (0.08)	0.89 (0.10)	1.37 (0.13)	0.44
159	Pick up and leave when tired	0.80 (0.15)	-0.95 (0.26)	0.70 (0.24)	1.84 (0.40)	0.31
279	Doesn't turn down a dare.	1.27 (0.19)	0.22(0.14)	1.39 (0.21)	2.33 (0.35)	0.35
				ANT-S-R		
39	Kick out of dangerous things	2.07 (0.26)	0.41 (0.09)	1.31 (0.14)	1.88 (0.21)	0.48
79	Do wild things for fun	3.27 (0.34)	0.19 (0.07)	1.00 (0.09)	1.52 (0.12)	0.50
119	Behaviour is wild.	2.41 (0.27)	-0.07 (0.09)	0.91 (0.10)	1.39 (0.14)	0.49
279	Doesn't turn down a dare.	1.21 (0.19)	0.23 (0.15)	1.43 (0.23)	2.41 (0.38)	0.37

[a]Following the removal of items #199, 239, and 313.

Table 13

Item parameters and scalability coefficients and estimates of NCDIF of MMT population for the ANX-A scale ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*	*NCDIF*
				ANX-A[a]			
4	Extremely tense	1.90 (0.19)	-0.49 (0.11)	0.62 (0.11)	1.33 (0.15)	0.47	**0.165**
44	Nervousness	2.15 (0.25)	0.14 (0.09)	0.94 (0.11)	1.69 (0.17)	0.47	0.004
84	Afraid for no reason	1.96 (0.25)	-0.09 (0.10)	0.87 (0.12)	1.40 (0.15)	0.46	0.010
204	Feel something terrible	1.76 (0.21)	-0.25 (0.11)	0.90 (0.13)	1.58 (0.19)	0.44	0.031
284	Easily startled	1.65 (0.21)	-0.09 (0.11)	1.04 (0.15)	1.73 (0.21)	0.43	**0.356**.
				ANX-A-R			
4	Extremely tense	1.67 (0.18)	-0.52 (0.12)	0.66 (0.12)	1.42 (0.17)	0.45	-
44	Nervousness	2.42 (0.28)	0.13 (0.08)	0.89 (0.10)	1.61 (0.15)	0.51	-
84	Afraid for no reason	2.03 (0.25)	-0.10 (0.09)	0.86 (0.12)	1.38 (0.15)	0.49	-
204	Feel something terrible	1.75 (0.21)	-0.25 (0.11)	0.91 (0.13)	1.58 (0.18)	0.45	-

[a]Following the removal of items #124, 164, and 244

Table 14

Item parameters and scalability coefficients for the ANX-C subscale and the ANX-C-R subscale ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	Item *H*
						ANX-C[a]				
25	Trouble concentrating, nervous	2.04	(0.21)	-0.35	(0.10)	0.54	(0.11)	1.36	(0.14)	0.53
65	Hard to enjoy because of worry	2.17	(0.22)	-0.87	(0.11)	0.16	(0.09)	0.82	(0.11)	0.56
105	So worried can't stand it	2.99	(0.29)	-0.05	(0.08)	0.85	(0.09)	1.43	(0.12)	0.58
145	Told I worry too much	1.88	(0.20)	-0.74	(0.12)	0.31	(0.11)	0.96	(0.13)	0.51
265	Worry more than necessary	2.01	(0.20)	-1.37	(0.15)	-0.18	(0.09)	0.62	(0.11)	0.54
305	So nervous, afraid will die	1.24	(0.20)	0.76	(0.16)	1.61	(0.25)	2.13	(0.33)	0.41
						ANX-C-R				
25	Trouble concentrating, nervous	2.18	(0.22)	-0.33	(0.09)	0.53	(0.10)	1.32	(0.14)	0.57
65	Hard to enjoy because of worry	2.15	(0.22)	-0.87	(0.11)	0.16	(0.09)	0.82	(0.12)	0.59
105	So worried can't stand it	2.97	(0.29)	-0.05	(0.07)	0.85	(0.09)	1.44	(0.13)	0.62
265	Worry more than necessary	1.92	(0.20)	-1.39	(0.15)	-0.18	(0.10)	0.63	(0.12)	0.56

[a]Following the removal of items #185 and 225

Table 15

Item parameters and scalability coefficients from the analysis of the ANX-P subscale in a MMT population ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
				ANX-P[a]		
33	Often jittery	1.47 (0.20)	-0.10 (0.12)	1.17 (0.17)	1.88 (0.24)	0.39
73	Worry so much, feels like fainting	2.12 (0.29)	0.60 (0.10)	1.39 (0.15)	2.06 (0.23)	0.43
113	Dizzy when under pressure	2.05 (0.24)	-0.07 (0.10)	0.92 (0.12)	1.46 (0.15)	0.44
153	Often feel heart pounding	1.55 (0.19)	-0.65 (0.14)	0.66 (0.13)	1.42 (0.18)	0.43
233	Trouble breathing under pressure	2.25 (0.28)	0.35 (0.09)	1.12 (0.12)	1.67 (0.17)	0.43
273	Sweating hands often	1.49 (0.21)	0.15 (0.12)	1.08 (0.16)	1.85 (0.24)	0.41
				ANX-P-R		
73	Worry so much, feels like fainting	2.21 (0.29)	0.60 (0.10)	1.38 (0.14)	2.04 (0.22)	0.46
113	Dizzy when under pressure	2.52 (0.28)	-0.06 (0.08)	0.86 (0.10)	1.36 (0.13)	0.51
153	Often feel heart pounding	1.27 (0.18)	-0.73 (0.16)	0.73 (0.16)	1.60 (0.24)	0.40
233	Trouble breathing under pressure	2.11 (0.26)	0.35 (0.09)	1.15 (0.12)	1.72 (0.18)	0.44

[a]Following the removal of items #193 and 313

Table 16

Item parameters and scalability coefficients from the analysis of the ARD- O-R subscale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	Item *H*
5	Do things a certain way	2.13	(0.24)	-0.30	(0.09)	0.58	(0.09)	1.32	(0.15)	0.48
45	Fight to control impulses	1.71	(0.22)	-0.06	(0.10)	0.79	(0.13)	1.59	(0.19)	0.52
85	Bothered by things out of place	1.07	(0.16)	-1.22	(0.20)	0.48	(0.16)	1.46	(0.25)	0.50

Table 17

Item parameters and scalability coefficients for the ARD- P-R subscale ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
26	Fear of saying something wrong	1.93 (0.22)	-0.33 (0.10)	0.80 (0.11)	1.59 (0.17)	0.48
66	Exaggerated fears	2.37 (0.28)	0.12 (0.08)	1.11 (0.12)	1.71 (0.15)	0.52
106	Nervous in front of others	1.67 (0.20)	-0.83 (0.13)	0.37 (0.11)	1.00 (0.14)	0.50

Table 18

Item parameter values, standard errors, and scalability coefficients of MMT population for the ARD-T scale ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	Item *H*
34	Reliving horrible event	3.25 (0.40)	0.14 (0.07)	0.78 (0.09)	1.19 (0.10)	0.61
74	Bothered by thoughts of past	2.12 (0.23)	-0.61 (0.11)	0.50 (0.10)	1.16 (0.13)	0.55
114	Memories of bad experience	3.08 (0.33)	-0.16 (0.08)	0.51 (0.08)	0.94 (0.10)	0.58
154	Can't get over something	3.18 (0.33)	-0.35(0.08)	0.39 (0.07)	0.88 (0.09)	0.60
194	Horrible experience causes guilt	1.95 (0.23)	-0.50 (0.11)	0.35 (0.11)	0.89 (0.13)	0.52
234	Nightmares about past	2.40 (0.27)	0.26 (0.09)	0.86 (0.11)	1.36 (0.14)	0.55
274	No longer interested in some things	1.94 (0.20)	-0.13 (0.10)	0.66 (0.11)	1.31 (0.16)	0.51
314	Avoid things that bring memories	1.13 (0.15)	-1.32 (0.22)	-0.21 (0.15)	0.91 (0.20)	0.42

Table 19

Item parameter values, standard errors, and scalability coefficients for the ARD-T-R scale in an MMT sample ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*
34	Reliving horrible event	2.99 (0.31)	0.16 (0.07)	0.81 (0.09)	1.23 (0.11)	0.66
114	Troubled by bad memories	3.33 (0.36)	-0.15 (0.07)	0.52 (0.08)	0.93 (0.09)	0.68
154	Can't get over something	3.30 (0.34)	-0.33 (0.07)	0.40 (0.07)	0.89 (0.09)	0.69
234	Nightmares about past	2.51 (0.27)	0.26 (0.08)	0.86 (0.10)	1.34 (0.14)	0.64

Table 20

Item parameter values, standard errors, and scalability coefficients for the BOR A subscale in a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
14	Sudden mood shifts	2.21	(0.24)	-0.97	(0.11)	0.22	(0.09)	0.78	(0.11)	0.53
54	Intense mood shifts	3.01	(0.31)	-0.47	(0.08)	0.58	(0.07)	1.24	(0.11)	0.54
214	Can't do enough to express anger	1.42	(0.17)	-0.68	(0.15)	0.42	(0.13)	1.15	(0.18)	0.44

Table 21

Item parameter values, standard errors, and scalability coefficients for the BOR-I subscale and the BOR-I-Rsubscale in a MMT sample(N = 323).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
				BOR-I[a]			
17	Attitude about self changes	1.71 (0.19)	-0.70 (0.13)	0.42 (0.11)	1.19 (0.16)	0.45	0.049
57	Feel empty inside	2.68 (0.26)	-0.77 (0.10)	0.30 (0.08)	0.80 (0.09)	0.51	0.208
97	Worry about abandonment	1.58 (0.19)	-0.37 (0.12)	0.50 (0.12)	1.27 (0.17)	0.45	0.040
137	Wonder what to do with life	2.06 (0.22)	-1.17 (0.13)	-0.01 (0.09)	0.60 (0.10)	0.49	0.028
177	Can't handle separation	1.40 (0.17)	-0.92 (0.16)	0.24 (0.13)	1.20 (0.19)	0.42	0.069
				BOR-I-R			
17	Attitude about self changes	1.58 (0.20)	-0.73 (0.13)	0.44 (0.15)	1.24 (0.15)	0.50	
57	Feel empty inside	3.13 (0.32)	-0.74 (0.09)	0.29 (0.07)	0.78 (0.08)	0.58	
137	Wonder what to do with life	2.09 (0.23)	-1.17 (0.12)	-0.02 (0.09)	0.60 (0.09)	0.57	

[a]Following the removal of item #217

Table 22

Item parameters, scalability coefficients, and differential item and test functioning coefficients for the remaining items of the PAI subscale BOR-N and the BOR-N-R based upon responses to the PAI by a sample of MMT patients(N = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*C-DIF*	*NC-DIF*
						BOR-N[a]						
19	Stormy relationships.	1.09	(0.17)	-0.83	(0.20)	0.59	(0.16)	1.60	(0.26)	0.31	0.004	0.026
59	Let people know they hurt me.	1.29	(0.18)	-1.34	(0.20)	0.08	(0.12)	0.94	(0.16)	0.35	0.014	0.047
99	People let me down.	2.85	(0.30)	-1.17	(0.11)	-0.10	(0.07)	0.43	(0.08)	0.44	0.027	0.080
179	Picked wrong friends.	1.45	(0.19)	-1.58	(0.22)	-0.33	(0.13)	0.45	(0.12)	0.38	-0.031	0.298
						BOR-N-R						
59	Let people know they hurt me.	1.46	(0.18)	-1.25	(0.18)	0.07	(0.11)	0.87	(0.14)	0.42	-	-
99	People let me down.	2.31	(0.25)	-1.25	(0.13)	-0.11	(0.08)	0.47	(0.09)	0.45	-	-
179	Picked wrong friends.	1.50	(0.19)	-1.55	(0.20)	-0.33	(0.12)	0.45	(0.12)	0.41	-	-

[a]Following the removal of items #139 and 219

Table 23

Item parameters, scalability coefficients, and differential item and test functioning coefficients for the remaining items of the PAI subscale BOR-S and the BOR-S-R based upon responses to the PAI by a sample of MMT patients ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*
				BOR-S[a]		
143	Impulsivity causes trouble	1.77 (0.21)	-0.24 (0.11)	0.79 (0.12)	1.32 (0.15)	0.41
183	Upset hurt self	1.29 (0.24)	1.22 (0.19)	1.99 (0.30)	2.59 (0.43)	0.32
223	Too impulsive	2.34 (0.25)	-0.03 (0.09)	0.83 (0.10)	1.48 (0.14)	0.44
303	Reckless	1.79 (0.25)	0.58 (0.11)	1.53 (0.18)	2.22 (0.26)	0.39
				BOR-S-R		
143	Impulsivity causes trouble	1.70 (0.21)	-0.25 (0.11)	0.79 (0.13)	1.34 (0.16)	0.44
223	Too impulsive	2.70 (0.30)	-0.03 (0.08)	0.79 (0.09)	1.41 (0.13)	0.48
303	Reckless	1.65 (0.24)	0.60 (0.11)	1.59 (0.20)	2.31 (0.30)	0.42

[a]Following the removal of items #263 and 343

Table 24

Item parameter values, standard errors, and scalability coefficients of the DEP-A subscale of the PAI when employed in a sample of MMT patients following the removal of items #246 and #248 ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF (Prob.)*
6	Often sad no reason	1.81 (0.22)	-0.17 (0.11)	0.82 (0.12)	1.63 (0.19)	0.43	0.002 (0.00)
46	Forgotten feeling happy	2.39 (0.26)	0.03 (0.09)	0.93 (0.10)	1.53 (0.14)	0.46	0.017 (0.00)
86	Everything is a big effort	1.96 (0.23)	-0.21 (0.10)	0.83 (0.12)	1.55 (0.17)	0.45	0.054 (0.00)
126	No pleasure	2.31 (0.27)	0.16 (0.09)	0.98 (0.11)	1.87 (0.19)	0.46	0.001 (0.10)
166	No interest in enjoyable things	1.25 (0.17)	-1.21 (0.20)	0.20 (0.13)	1.05 (0.18)	0.39	0.003 (0.00)
206	No interest in life	1.93 (0.31)	0.99 (0.13)	1.72 (0.21)	2.36 (0.33)	0.41	0.001 (0.43)

Table 25

Item parameters and scalability coefficients of the DEP-A-R subscale based upon responses of a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
6	Often sad no reason	1.71	(0.21)	-0.17	(0.11)	0.84	(0.12)	1.69	(0.20)	0.44
46	Forgotten feeling happy	2.45	(0.27)	0.03	(0.09)	0.94	(0.10)	1.54	(0.14)	0.47
126	No pleasure	2.43	(0.28)	0.15	(0.09)	0.97	(0.10)	1.85	(0.17)	0.49
206	No interest in life	1.77	(0.29)	1.03	(0.14)	1.79	(0.23)	2.48	(0.35)	0.41

Table 26

Item means and scalability coefficients of the DEP-C subscale based upon responses of a sample of MMT patients ($N = 323$).

Item#	Content	Mean	*Item H*
27	Let everyone down	1.36	0.28
67	Worthless	0.98	0.37
107	Don't feel like trying	0.66	0.37
147	Can't concentrate	1.04	0.28
187	Nothing works	0.67	0.36
227	Good things will happen	1.17	0.16
267	Have something to contribute	1.50	0.14
307	Pretty successful	1.51	0.14

Table 27

Item parameters and scalability coefficients of the remaining items of the DEP-C and DEP-C-R subscales based upon responses of a sample of MMT patients ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*
				DEP-C[a]		
27	Let everyone down	1.86 (0.21)	-0.71 (0.12)	0.25 (0.10)	0.95 (0.14)	0.49
67	Worthless	2.98 (0.33)	-0.18 (0.07)	0.75 (0.09)	1.19 (0.11)	0.55
107	Don't feel like trying	2.32 (0.28)	0.26 (0.09)	1.26 (0.13)	1.77 (0.18)	0.49
147	Can't concentrate	1.37 (0.19)	-0.41 (0.14)	0.89 (0.16)	1.74 (0.24)	0.42
187	Nothing works	2.25 (0.27)	0.20 (0.09)	1.29 (0.13)	1.77 (0.19)	0.50
				DEP-C-R		
27	Let everyone down	1.89 (0.21)	-0.71 (0.12)	0.25 (0.10)	0.95 (0.13)	0.53
67	Worthless	3.43 (0.34)	-0.18 (0.07)	0.72 (0.08)	1.15 (0.09)	0.60
107	Don't feel like trying	2.20 (0.26)	0.26 (0.09)	1.29 (0.13)	1.82 (0.19)	0.52
187	Nothing works	2.00 (0.25)	0.21 (0.10)	1.36 (0.15)	1.88 (0.21)	0.51

[a]Following the elimination of items #227, 267, and 307

Table 28

Item parameters and scalability coefficients of the DEP-P-R subscale based upon responses of a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
35	Hardly any energy	1.40	(0.19)	-0.35	(0.12)	1.02	(0.16)	1.99	(0.26)	0.40
155	Moving more slowly	1.26	(0.19)	-0.08	(0.13)	1.08	(0.18)	2.04	(0.29)	0.38
195	Wake up early, can't sleep	1.45	(0.18)	-0.92	(0.15)	0.16	(0.12)	0.85	(0.14)	0.39
275	Often wake up in night	2.16	(0.24)	-0.93	(0.12)	-0.09	(0.09)	0.49	(0.09)	0.47

Table 29

Item parameter values, standard errors, and scalability coefficients of the DOM-R subscale of the PAI when employed in a sample of MMT patients following the removal of items #246 and #248 ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*	*p*
16	Take charge person	2.10	(0.22)	-1.03	(0.12)	0.07	(0.09)	1.03	(0.12)	0.50	0.013	0.00
56	Natural leader	2.12	(0.21)	-0.48	(0.10)	0.47	(0.10)	1.31	(0.14)	0.48	0.022	0.00
96	Good at telling others	1.56	(0.19)	-0.69	(0.13)	0.48	(0.12)	1.33	(0.18)	0.46	0.007	0.36
176	Best as a leader	1.73	(0.20)	-0.47	(0.11)	0.77	(0.12)	1.52	(0.18)	0.48	0.016	0.00

Table 30

Item parameter values, standard errors, and scalability coefficients for the DRG-R subscale in a MMT patient sample (N = 323).

Item	Content	a	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
22	Use to feel better	1.50	(0.18)	-1.84	(0.24)	-0.67	(0.13)	0.24	(0.12)	0.33
62	Told have a drug problem	1.65	(0.19)	-1.77	(0.21)	-0.75	(0.12)	0.12	(0.11)	0.36
102	Caused financial strain	1.46	(0.18)	-1.91	(0.26)	-0.76	(0.14)	-0.17	(0.12)	0.34
222	Use is out of control	1.46	(0.18)	-0.45	(0.13)	0.43	(0.12)	0.98	(0.16)	0.37
262	Caused health problems	1.28	(0.16)	-0.85	(0.16)	0.18	(0.14)	0.87	(0.17)	0.33

Table 31

Item parameter values, standard errors, and scalability coefficients of the MAN-A-R subscale of the PAI when employed in a sample of MMT patients following the removal of items #246 and #248 ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*
47	Take on too many commitments	1.34 (0.17)	0.07 (0.12)	1.51 (0.21)	2.26 (0.31)	0.34
127	Racing thoughts	1.39 (0.16)	-1.39 (0.18)	0.31 (0.12)	1.18 (0.17)	0.37
207	Need to keep active, not rest	1.49 (0.19)	-0.17 (0.11)	0.94 (0.14)	1.93 (0.24)	0.37

Table 32

Item parameters, scalability coefficients, and estimates of NCDIF (MAN-G and MAN-G-R subscalea) following the removal of items that did not conform to the assumption of monotone homogeneity ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*	*p*
						MAN-G[a]						
28	Many brilliant ideas	1.23	(0.17)	-1.47	(0.21)	0.42	(0.14)	1.71	(0.25)	0.32	0.065	0.000
68	Very special talents	1.54	(0.19)	-0.57	(0.13)	0.61	(0.12)	1.46	(0.19)	0.35	0.087	0.000
108	Plans will make me famous	1.96	(0.25)	0.36	(0.09)	1.27	(0.14)	1.91	(0.19)	0.38	0.034	0.000
188	Answers to important questions	1.22	(0.16)	-0.82	(0.17)	0.76	(0.15)	1.70	(0.24)	0.31	0.154	0.000
						MAN-G-R						
28	Many brilliant ideas	1.42	(0.16)	-1.34	(0.18)	0.40	(0.12)	1.57	(0.22)	0.37	-	-
68	Very special talents	1.77	(0.21)	-0.54	(0.12)	0.57	(0.10)	1.36	(0.17)	0.39	-	-
108	Plans will make me famous	1.42	(0.21)	0.41	(0.12)	1.49	(0.21)	2.26	(0.30)	0.36	-	-

[a]Not including item #148, 228, 268, and 308

Table 33

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the MAN-I subscale in a sample of MMT patients (N = 323)

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*	*p*
36	Very demanding	1.30	(0.17)	-1.48	(0.22)	-0.02	(0.13)	0.90	(0.17)	0.36	0.004	0.4197
76	Irritated when kept from goals	1.29	(0.16)	-1.69	(0.23)	-0.18	(0.13)	0.87	(0.17)	0.38	0.073	0.0000
116	Upset when others don't understand	1.74	(0.20)	-0.57	(0.12)	0.81	(0.12)	1.65	(0.19)	0.42	0.003	0.0000
156	Irritated when plans are interfered	1.35	(0.18)	0.06	(0.13)	1.26	(0.18)	1.90	(0.26)	0.37	0.025	0.0000
196	Bothered by not being understood	1.88	(0.20)	-0.62	(0.11)	0.62	(0.11)	1.33	(0.15)	0.42	0.014	0.0000
236	No patience with being held back	1.67	(0.18)	-0.97	(0.15)	-0.02	(0.11)	0.79	(0.13)	0.41	0.005	0.4553
276	Very touchy, easily annoyed	1.35	(0.17)	-1.22	(0.20)	0.21	(0.13)	1.02	(0.17)	0.35	0.022	0.0000
316	No patience for those who disagree	1.50	(0.18)	-0.48	(0.13)	0.77	(0.14)	1.52	(0.20)	0.37	0.003	0.0000

Table 34

Item parameter values, standard errors, and scalability coefficients for the MAN-I-R subscale in a MMT sample (N = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
116	Upset when others don't understand	1.28	(0.17)	-1.70	(0.23)	-0.18	(0.13)	0.86	(0.17)	0.42
196	Bothered by those who don't understand	1.67	(0.20)	-0.65	(0.12)	0.65	(0.12)	1.39	(0.18)	0.47
236	No patience with being held back	2.17	(0.23)	-0.88	(0.11)	-0.04	(0.09)	0.70	(0.10)	0.43
316	No patience for those who disagree	1.48	(0.19)	-0.49	(0.13)	0.77	(0.14)	1.52	(0.21)	0.43

Table 35

Item parameters and scalability coefficients of the NIM-R subscale based upon responses to the PAI by a sample of MMT patients (*N* = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
9	Cannot remember who I am.	1.93	(0.42)	1.82	(0.22)	2.56	(0.34)	3.26	(0.55)	0.40
89	Destined to unhappiness	1.79	(0.24)	0.73	(0.11)	1.48	(0.17)	2.12	(0.25)	0.37
129	Multiple personalities	1.39	(0.21)	0.66	(0.14)	1.82	(0.25)	2.69	(0.37)	0.34
209	Lose memory	1.79	(0.24)	0.58	(0.11)	1.37	(0.16)	1.97	(0.22)	0.39
289	No happy childhood memories	1.91	(0.25)	0.49	(0.10)	1.15	(0.14)	1.78	(0.20)	0.41

Table 36

Item parameter values, standard errors, and scalability coefficients for the PIM scale following the removal of an item lacking monotonicity ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
24	little things bother me	1.77 (0.20)	-0.68 (0.12)	0.24 (0.11)	1.32 (0.16)	0.41	0.015
64	avoid someone I don't like	1.24 (0.17)	-0.22 (0.14)	0.85 (0.16)	2.20 (0.32)	0.35	0.001
104	complain too much	1.36 (0.17)	-1.60 (0.23)	-0.84 (0.16)	0.74 (0.15)	0.36	0.209*
144	too impatient	1.79 (0.19)	-0.68 (0.12)	0.14 (0.11)	1.23 (0.16)	0.40	0.008
184	don't take criticism	1.21 (0.16)	-1.75 (0.26)	-0.75 (0.17)	1.06 (0.20)	0.35	0.010
224	put things off	2.12 (0.22)	-0.45 (0.10)	0.29 (0.09)	1.34 (0.15)	0.44	0.038
264	make promises I can't keep	1.44 (0.18)	-1.23 (0.18)	-0.34 (0.12)	1.20 (0.18)	0.38	0.172*
304	could have been more thoughtful	1.13 (0.16)	-0.42 (0.16)	0.35 (0.16)	1.96 (0.32)	0.30	0.005

* above the 0.054 cut off and statistically significant at the .05 level.

Table 37

Item parameters and scalability coefficients of the NON-R subscale based upon responses to the PAI by a sample of MMT patients (*N* = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
1	Friends available	1.10	(0.12)	-1.76	(0.27)	-0.46	(0.16)	1.16	(0.22)	0.35
81	People to talk to	2.02	(0.20)	-0.70	(0.11)	0.16	(0.09)	1.24	(0.13)	0.44
161	People supportive	1.80	(0.19)	-0.61	(0.11)	0.38	(0.11)	1.45	(0.16)	0.40
201	People care about me	1.85	(0.19)	-0.46	(0.11)	0.52	(0.11)	1.84	(0.21)	0.43

Table 38

Item parameter values, standard errors, and scalability coefficients for the final set of the PIM-R scale ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*
24	little things bother me	1.87	(0.20)	-0.66	(0.11)	0.24	(0.10)	1.29	(0.13)
104	complain too much	1.45	(0.17)	-1.54	(0.20)	-0.81	(0.14)	0.72	(0.13)
144	too impatient	1.65	(0.18)	-0.71	(0.13)	0.14	(0.11)	1.28	(0.15)
224	put things off	2.12	(0.21)	-0.44	(0.10)	0.29	(0.09)	1.35	(0.14)
264	make promises I can't keep	1.40	(0.17)	-1.25	(0.18)	-0.35	(0.13)	1.22	(0.17)

Table 39

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the PAR-H-R subscale in a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*	*p*
48	Alert to unfaithful people	1.39	(0.19)	-0.54	(0.14)	0.83	(0.15)	1.60	(0.23)	0.36	0.687	0.000
168	People hide real motives	1.28	(0.18)	-1.50	(0.23)	0.13	(0.11)	1.22	(0.21)	0.36	0.657	0.000

Table 40

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the PAR-P scale following the removal of three items lacking monotonicity in a sample of MMT patients(N = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*
						PAR-P[a]					
29	People go out of way to bother me	1.59	(0.21)	0.14	(0.11)	1.47	(0.18)	2.03	(0.25)	0.39	0.01
69	People make me look bad	2.36	(0.27)	0.13	(0.09)	1.01	(0.11)	1.46	(0.13)	0.46	0.01
149	People try to keep me from getting ahead.	1.77	(0.21)	-0.03	(0.11)	1.10	(0.14)	1.69	(0.19)	0.42	0.00
189	People keep me from getting ahead	1.71	(0.24)	0.41	(0.11)	1.30	(0.16)	1.97	(0.24)	0.39	0.01
269	People have it in for me.	1.79	(0.23)	0.29	(0.10)	1.38	(0.16)	2.11	(0.25)	0.43	0.00
						PAR-P-R					
69	People make me look bad	2.15	(0.26)	0.14	(0.09)	1.05	(0.12)	1.52	(0.15)	0.46	
149	Some people try to keep me from getting ahead.	1.74	(0.21)	-0.03	(0.11)	1.11	(0.14)	1.70	(0.19)	0.43	
189	People keep me from getting ahead	1.91	(0.25)	0.40	(0.10)	1.24	(0.14)	1.87	(0.21)	0.42	
269	People have it in for me.	1.85	(0.23)	0.29	(0.10)	1.37	(0.15)	2.08	(0.23)	0.44	

[a]Not including eliminated items #109, 229, and 309.

Table 41

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the PAR-R-R subscale in a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*
117	Given a lot, not much in return	2.03	(0.23)	-0.66	(0.10)	0.16	(0.08)	1.05	(0.12)	0.53	0.001
157	People don't appreciate me	2.24	(0.25)	-0.64	(0.10)	0.43	(0.09)	1.09	(0.13)	0.53	0.005

Table 42

Item parameters, scalability coefficients, and NCDIF for the final four retained items of the RXR-R subscale for use in MMT populations ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
2	Inner struggles cause problems.	1.27 (0.18)	-1.22 (0.19)	-0.01 (0.13)	1.58 (0.22)	0.39	0.181
42	Need to make changes.	1.43 (0.23)	0.80 (0.14)	1.86 (0.24)	2.95 (0.43)	0.44	0.321
82	Need to change even if it hurts.	1.60 (0.20)	-0.01 (0.11)	0.90 (0.14)	1.97 (0.24)	0.41	0.007
122	Need help to deal with problems.	2.24 (0.24)	-0.70 (0.10)	-0.05 (0.08)	0.95 (0.11)	0.49	0.169

Table 43

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the SCZ-P-R scale in a sample of MMT patients ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
10	Ideas others think strange.	1.60 (0.21)	0.01 (0.11)	1.41 (0.17)	2.04 (0.25)	0.35	0.313*
50	People put thoughts into my head.	2.08 (0.26)	0.36 (0.09)	1.49 (0.15)	2.15 (0.21)	0.39	0.035
170	Heard voices no one else heard.	1.35 (0.24)	1.31 (0.20)	1.98 (0.29)	2.54 (0.40)	0.34	0.005
210	People try to control my thoughts.	1.28 (0.21)	0.91 (0.17)	1.75 (0.26)	2.60 (0.40)	0.32	0.028

Table 44

Item parameters and scalability coefficients for the two reatained items of the PAI subscale SCZ-S-R ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*H*
30	Don't relate to people	1.70	(0.24)	0.24	(0.12)	1.60	(0.20)	2.44	(0.31)	0.38
70	Don't have much to say	1.34	(0.19)	-0.66	(0.16)	0.89	(0.17)	1.89	(0.27)	0.38

Table 45

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the SCZ T scale in a sample of MMT patients (N = 323) following the removal of item #318.

Item	Content	*a* SE	*b1* SE	*b2* SE	*b3* SE	*Item H*	*NCDIF*
38	confused thinking	1.99 (0.24)	-0.08 (0.10)	1.04 (0.12)	1.86 (0.20)	0.43	0.010
78	scrambled thoughts	2.04 (0.21)	-0.71 (0.12)	0.67 (0.11)	1.34 (0.14)	0.44	0.027
118	trouble keeping thoughts separate.	2.45 (0.28)	-0.06 (0.08)	0.96 (0.11)	1.71 (0.16)	0.47	0.002
158	somebody blocking thoughts.	1.46 (0.24)	0.64 (0.13)	1.42 (0.20)	2.25 (0.34)	0.39	0.036
198	shifting thoughts	1.72(0.20)	-0.72 (0.13)	0.53 (0.11)	1.28 (0.16)	0.44	0.074*
238	thoughts are being taken away.	1.39 (0.23)	0.83 (0.15)	1.85 (0.26)	2.65 (0.40)	0.40	0.024
278	thoughts disappear	1.81 (0.23)	-0.12 (0.11)	1.10 (0.14)	1.71 (0.20)	0.44	0.001

* Value exceeds cut off score of 0.54 and is statistically significant at the 0.05 level.

Table 46

Item parameter values, standard errors and scalability coefficients for the SCZ T-R scale (N = 323) following the removal of item #318.

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
38	confused thinking	2.05	(0.25)	-0.08	(0.10)	1.03	(0.12)	1.84	(0.19)	0.49
78	scrambled thoughts	2.08	(0.21)	-0.71	(0.11)	0.67	(0.11)	1.33	(0.14)	0.53
118	trouble keeping thoughts separate.	2.44	(0.28)	-0.06	(0.08)	0.97	(0.11)	1.71	(0.16)	0.52
278	thoughts disappear	1.79	(0.22)	-0.12	(0.11)	1.11	(0.14)	1.72	(0.20)	0.42

Table 47

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the SOM-C subscale following the removal of item #3 in a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*
43	Unexplainable illness	1.11	(0.21)	1.07	(0.20)	1.95	(0.33)	2.76	(0.48)	0.30	0.009
83	Numbness unexplained	2.13	(0.31)	0.92	(0.12)	1.58	(0.17)	1.92	(0.20)	0.45	0.042
123	Double or blurred vision	1.96	(0.24)	-0.05	(0.09)	0.76	(0.12)	1.23	(0.14)	0.45	0.035
163	Eyesight got worse then better	2.19	(0.26)	0.25	(0.09)	1.04	(0.12)	1.75	(0.18)	0.41	0.024
203	Lost feeling in hands	1.59	(0.21)	0.06	(0.11)	0.94	(0.14)	1.58	(0.20)	0.44	0.001
243	Legs so weak can't walk	2.19	(0.28)	0.46	(0.09)	1.12	(0.13)	1.45	(0.15)	0.41	0.003
283	Paralyzed	1.66	(0.23)	0.29	(0.11)	1.21	(0.16)	1.70	(0.22)	0.46	0.001

Table 48

Item parameter values and standard errors, scalability coefficients for the SOM-C-R subscale in a sample of MMT patients ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*
83	Numbness unexplained	1.76	(0.22)	-0.04	(0.10)	0.80	(0.13)	1.29	(0.16)	0.45
123	Double or blurred vision	3.27	(0.37)	0.23	(0.07)	0.93	(0.08)	1.56	(0.12)	0.49
163	Eyesight got worse then better	1.64	(0.21)	0.08	(0.11)	0.94	(0.14)	1.56	(0.19)	0.42
243	Legs so weak can't walk	1.42	(0.20)	0.33	(0.12)	1.33	(0.19)	1.86	(0.26)	0.39

Table 49

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the final four retained items of the SOM-H subscale for use in MMT populations ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
12	Seen a lot of doctors	1.26 (0.17)	-0.47 (0.15)	0.87 (0.16)	1.72 (0.25)	0.38	0.036
52	Complicated health problems	3.84 (0.44)	0.29 (0.07)	0.86 (0.07)	1.39 (0.11)	0.55	0.012
92	Health problems cause struggle	2.45 (0.28)	-0.05 (0.09)	0.75 (0.09)	1.24 (0.12)	0.49	0.026
132	Hard to treat health problems	3.13 (0.38)	0.46 (0.08)	0.99 (0.09)	1.6 (0.14)	0.53	0.005
212	Some unusual problems	1.63 (0.26)	0.84 (0.13)	1.53 (0.20)	1.90 (0.25)	0.42	0.019
252	Health is good	1.28 (0.17)	-1.21 (0.20)	-0.16 (0.14)	0.89 (0.17)	0.41	0.177*

* Value exceeds cut off score of 0.54 and is statistically significant at the 0.05 level.

Table 50

Item parameters for the final four retained items of the SOM-H-R subscale for use in MMT populations ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*H*
52	Complicated health problems	3.75	(0.44)	0.28	(0.07)	0.85	(0.07)	1.39	(0.11)	0.60
92	Medical problems cause struggle	2.44	(0.27)	-0.05	(0.09)	0.74	(0.09)	1.24	(0.13)	0.55
132	Difficult to treat medical issues	3.23	(0.39)	0.46	(0.07)	0.98	(0.09)	1.69	(0.14)	0.58
212	Unusual diseases and illnesses	1.57	(0.26)	0.86	(0.13)	1.57	(0.21)	1.95	(0.26)	0.44

Table 51

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the SOM-S-R subscale in a sample of MMT patients ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
32	Often don't feel well	1.95 (0.23)	-0.33 (0.10)	0.72 (0.11)	1.59 (0.17)	0.43	0.008
72	Suffer from pain	2.31(0.26)	-0.04 (0.09)	0.78 (0.09)	1.37 (0.13)	0.44	0.001
112	Good health	1.32(0.15)	-1.40 (0.20)	-0.31 (0.13)	0.94 (0.16)	0.37	0.240*
192	Bad back	1.03 (0.15)	0.04 (0.15)	1.10 (0.21)	1.55 (0.27)	0.32	0.142*

* Value exceeds cut off score of 0.54 and is statistically significant at the 0.05 level.

Table 52

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the STR scale in a sample of MMT patients ($N = 323$), following the removal of Item #326.

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*	*NCDIF*
321	Money problems.	2.18 (0.22)	-1.12 (0.12)	-0.13 (0.09)	0.52 (0.11)	0.46	0.195*
322	Life is unpredictable.	2.05 (0.21)	-0.84 (0.11)	0.20 (0.09)	0.91 (0.13)	0.44	0.079*
323	Many life changes.	1.52 (0.18)	-1.26 (0.18)	-0.38 (0.12)	0.61 (0.14)	0.39	0.305*
324	Unstable home.	1.45 (0.18)	-0.10 (0.12)	0.87 (0.15)	1.68 (0.24)	0.40	0.187*
325	Family problems	1.57 (0.20)	-0.17 (0.11)	0.88 (0.14)	1.55 (0.21)	0.42	0.160*
327	Money worries.	1.33 (0.18)	-1.90 (0.26)	-0.75 (0.15)	0.13 (0.13)	0.40	0.089*
328	Relationship problems.	1.06 (0.17)	-0.07 (0.15)	0.90 (0.20)	1.48 (0.28)	0.33	0.016

* Value exceeds cut off score of 0.54 and is statistically significant at the 0.05 level.

Table 53

Item parameters and scalability coefficients for the STR-R scale based upon response of a MMT sample ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*H*
321	Money problems.	1.98	(0.20)	-1.16	(0.13)	-0.13	(0.09)	0.55	(0.10)	0.47
322	Life is unpredictable.	2.47	(0.25)	-0.79	(0.09)	0.20	(0.08)	0.87	(0.10)	0.51
323	Many life changes.	1.51	(0.18)	-1.27	(0.17)	-0.39	(0.12)	0.62	(0.13)	0.43
325	Family problems	1.38	(0.18)	-0.17	(0.13)	0.95	(0.14)	1.67	(0.21)	0.43

Table 54

Item parameters for the SUI scale following the removal of Item #341 and Item # 301 ($N = 323$).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*
20	At times I wish I were dead.	2.48	(0.31)	0.03	(0.08)	0.96	(0.13)	1.34	(0.18)
60	thought about ways to kill self	4.15	(0.76)	0.22	(0.06)	0.80	(0.09)	1.05	(0.10)
100	made plans how to kill self	4.43	(1.54)	0.47	(0.06)	0.95	(0.09)	1.24	(0.12)
140	Recently thinking about suicide	3.86	(0.52)	0.61	(0.07)	1.05	(0.10)	1.40	(0.15)
180	Thought about suicide long time	5.06	(1.23)	0.45	(0.06)	0.93	(0.08)	1.25	(0.12)
220	Death would be relief	2.94	(0.39)	0.48	(0.09)	1.11	(0.13)	1.53	(0.17)
260	Thought about suicide note	3.00	(0.36)	0.44	(0.08)	0.95	(0.10)	1.38	(0.14)
261	No reasons to live	2.03	(0.30)	0.90	(0.15)	1.50	(0.24)	1.92	(0.31)
300	Thought about reaction of others	1.91	(0.25)	-0.19	(0.10)	0.45	(0.12)	0.89	(0.14)
340	Considering suicide	2.30	(0.37)	1.08	(0.14)	1.66	(0.23)	1.92	(0.27)

Table 55

Item parameters for the final six items retained for the SUI-R scale of the PAI ($N = 323$).

Item	Content	*a* *SE*	*b1* *SE*	*b2* *SE*	*b3* *SE*	*Item H*
60	thought about ways to kill self.	4.01 (0.50)	0.30 (0.06)	0.90 (0.09)	1.15 (0.11)	0.67
100	made plans how to kill self	4.57 (0.68)	0.55 (0.06)	1.04 (0.09)	1.32 (0.11)	0.65
140	Recently thinking about suicide	3.68 (0.62)	0.71 (0.08)	1.16 (0.10)	1.51 (0.16)	0.64
180	Thought about suicide long time	4.83 (0.68)	0.54 (0.07)	1.03 (0.08)	1.35 (0.13)	0.67
220	Death would be a relief.	2.68 (0.39)	0.58 (0.09)	1.22 (0.14)	1.67 (0.19)	0.59
260	Thought about suicide note	2.96 (0.39)	0.53 (0.09)	1.04 (0.10)	1.47 (0.15)	0.59

Table 56

Item parameter values, standard errors, scalability coefficients, and measures of DIF for the WRM-R subscale based upon responses to the PAI by a sample of MMT patients (N = 323).

Item	Content	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*Item H*	*NCDIF*
13	Very sociable	1.67	(0.17)	-1.68	(0.19)	-0.41	(0.11)	0.58	(0.12)	0.42	0.017
53	Making friends is easy	1.70	(0.21)	-1.26	(0.15)	-0.37	(0.11)	0.68	(0.13)	0.41	0.023
93	Like to meet new people	2.14	(0.22)	-1.42	(0.14)	-0.29	(0.09)	0.57	(0.10)	0.45	0.023
133	Warm person	1.26	(0.17)	-2.65	(0.39)	-1.01	(0.17)	0.44	(0.15)	0.37	0.005

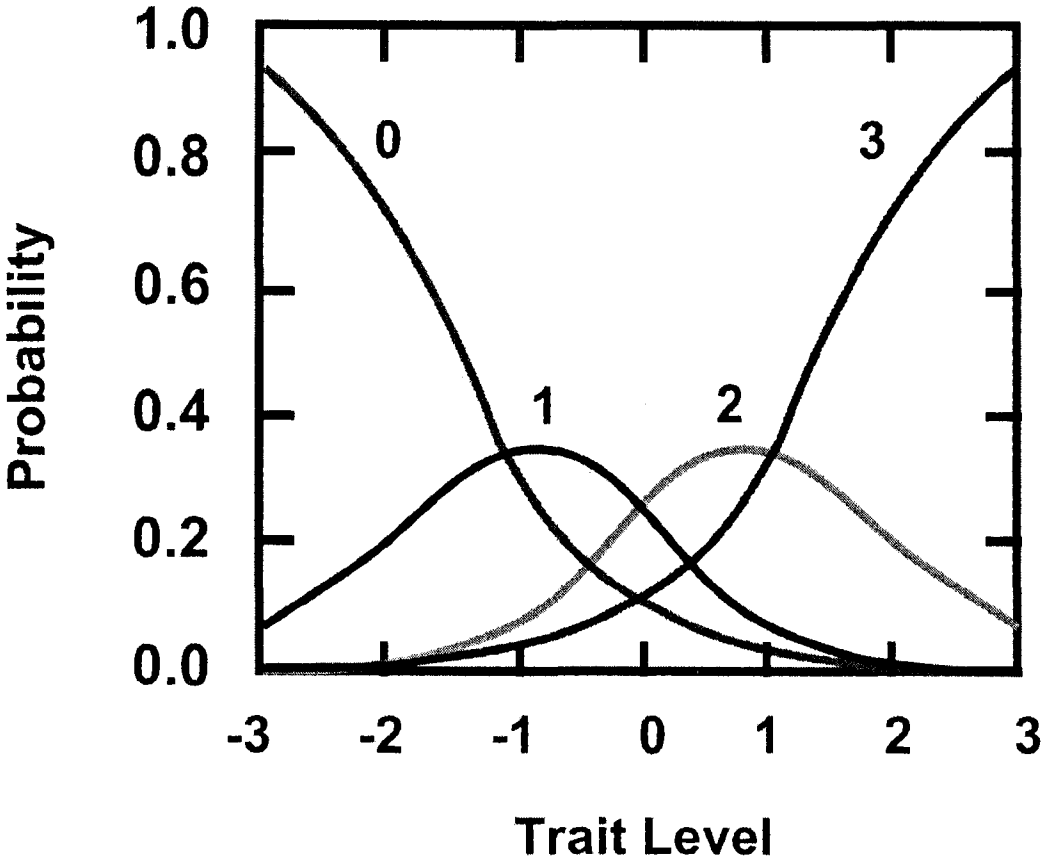


Figure 1. A typical Option Characteristic Curve for a multiple choice item, with four options, in which the item is effective across the entire range of the underlying trait. As can be seen individual's who score the lowest on the scale are most likely to choose option 0, while those who score the highest are most likely to choose option 3, with options 1 and 2 falling in the middle in an appropriate manner.

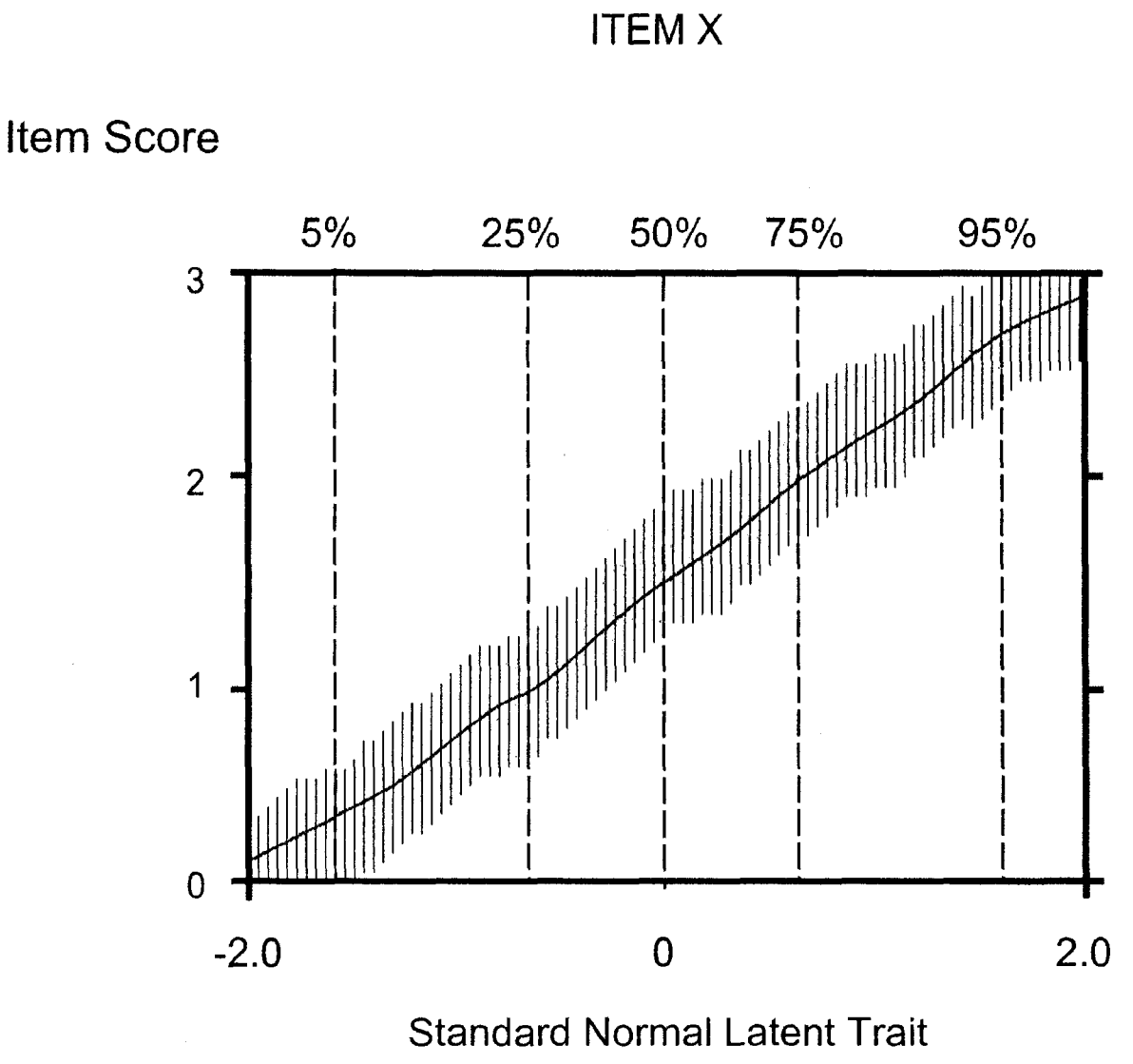
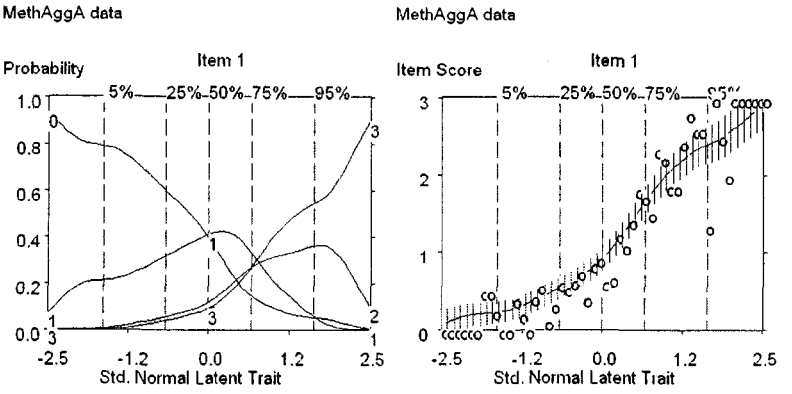
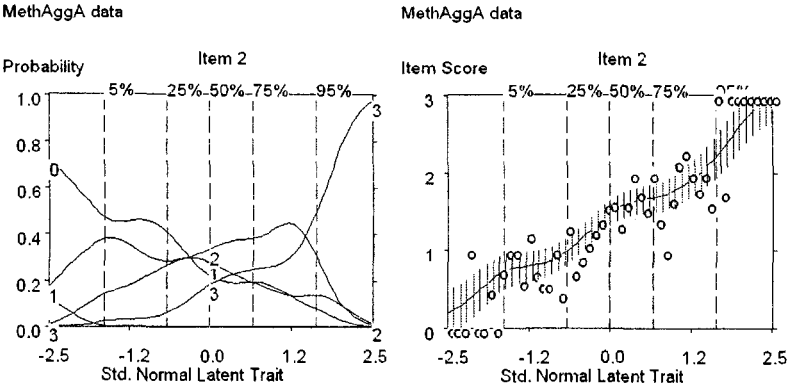


Figure 2. Item characteristic curve of a typical test item with four possible item responses.

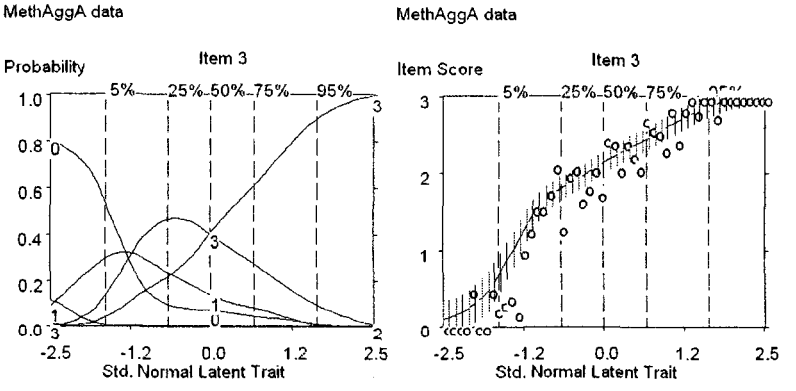
PAI Item # 258



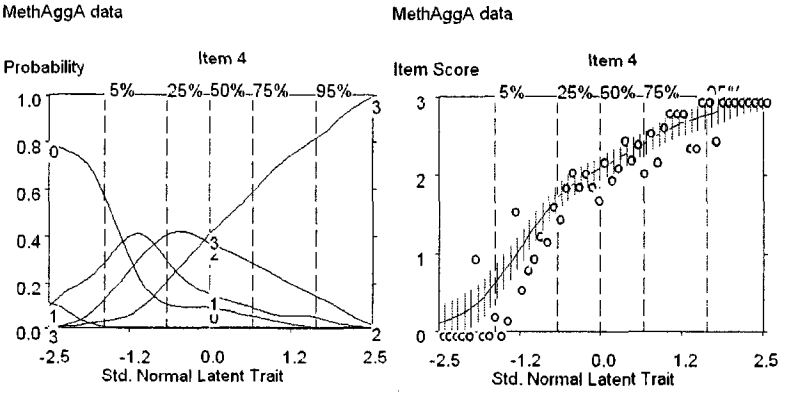
PAI Item # 259



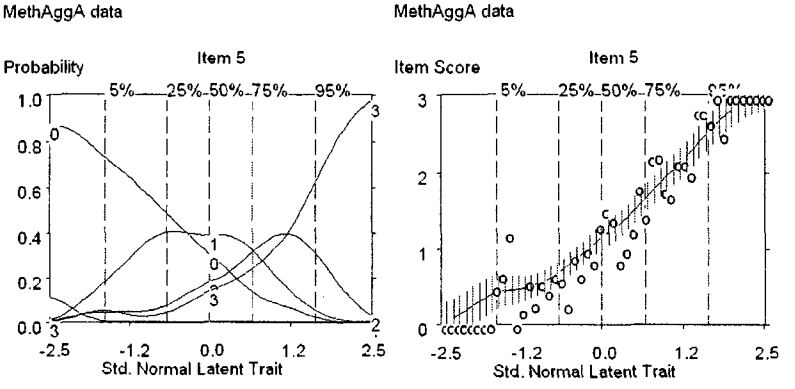
PAI Item # 298



PAI Item # 299



PAI Item # 338



PAI Item # 339

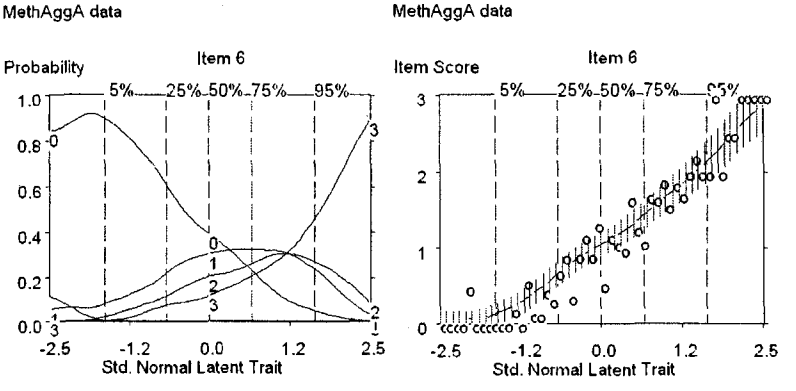


Figure 3. Option and item characteristic curves as calculated by Testgraf for the AGG-A scale administered to a sample of MMT patients ($N = 323$).

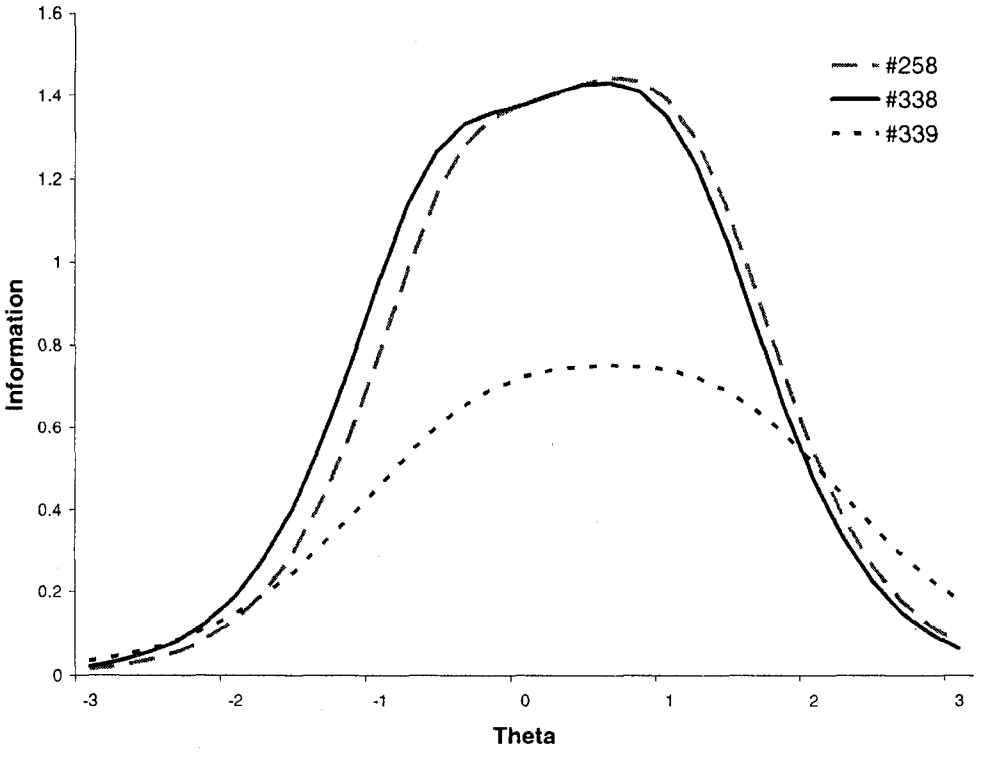


Figure 4. Information curves for the three items of the new AGG-A-R when administered to a sample of MMT patients ($N = 323$).

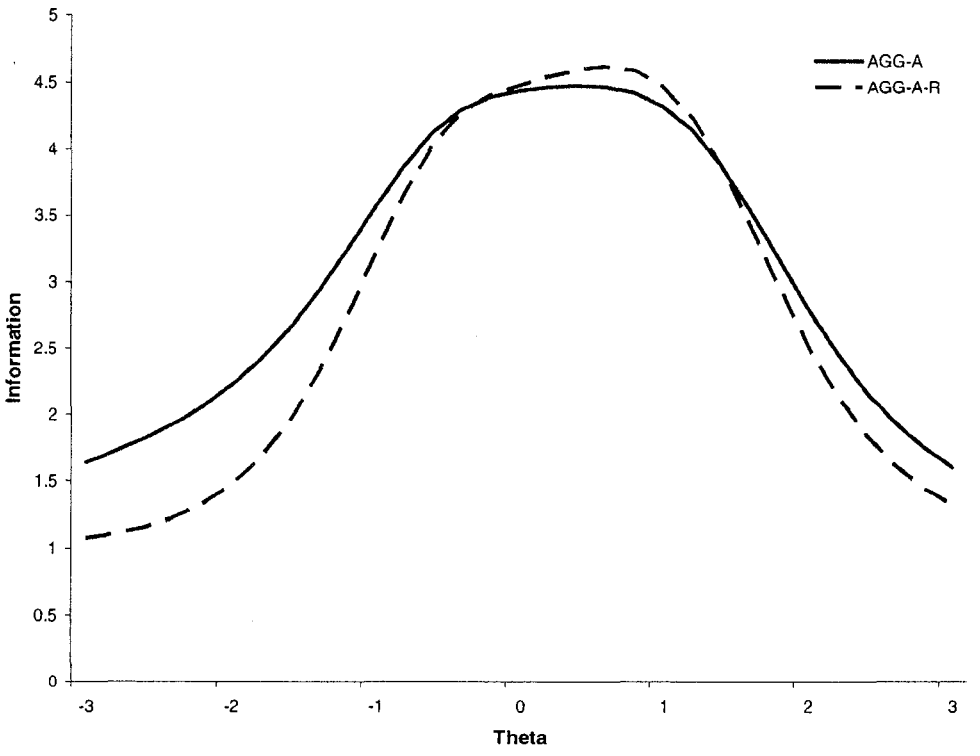
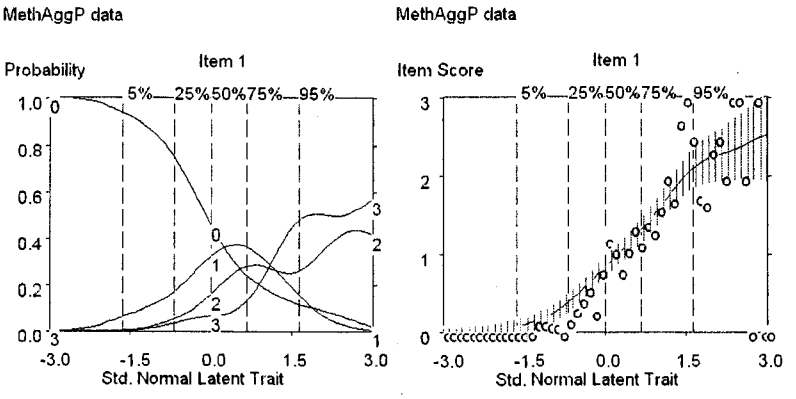
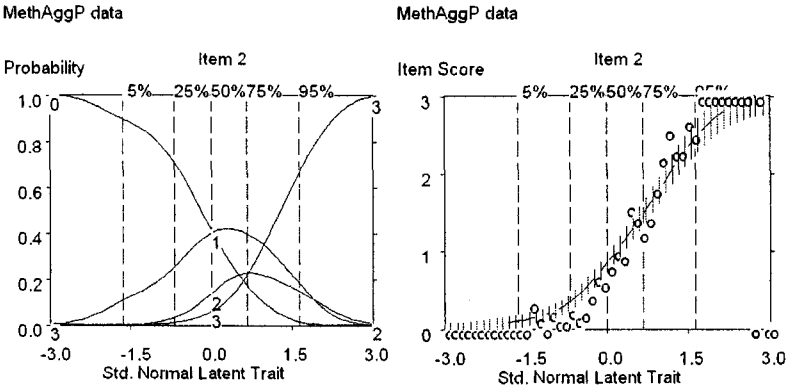


Figure 5. Comparison of total information of the original AGG-A subscale (6 items) and the AGG-A-R subscale (3 items) when employed in a sample of MMT patients (N = 323).

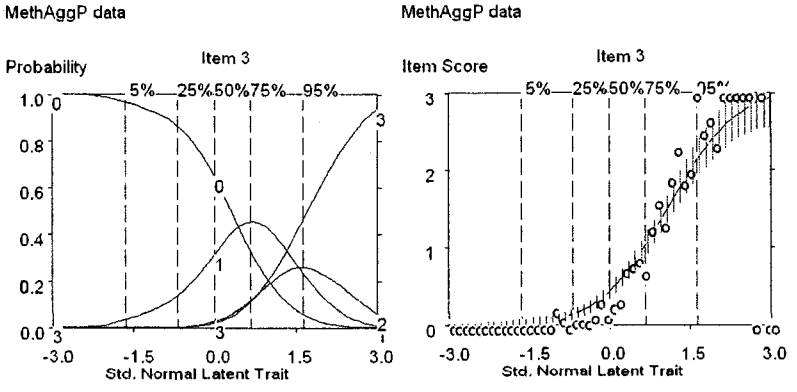
PAI Item #21



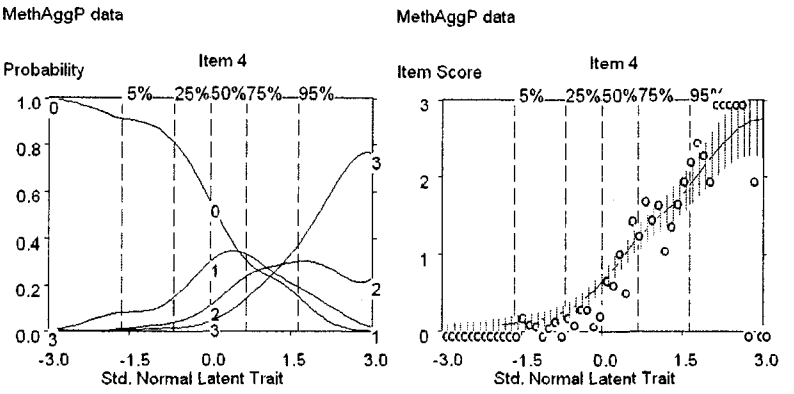
PAI Item # 61



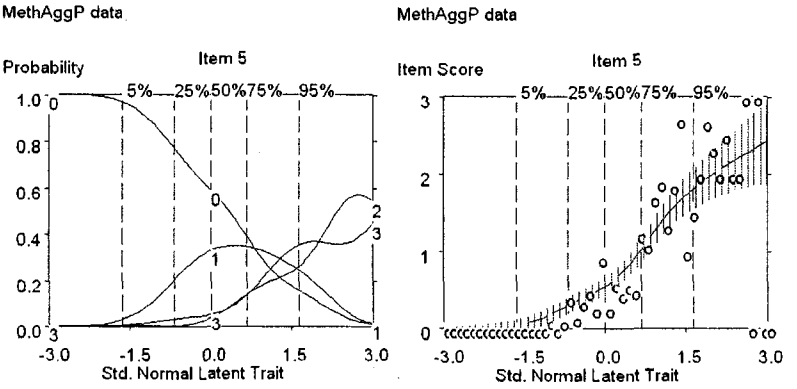
PAI Item #101



PAI Item #141



PAI Item #181



PAI Item # 221

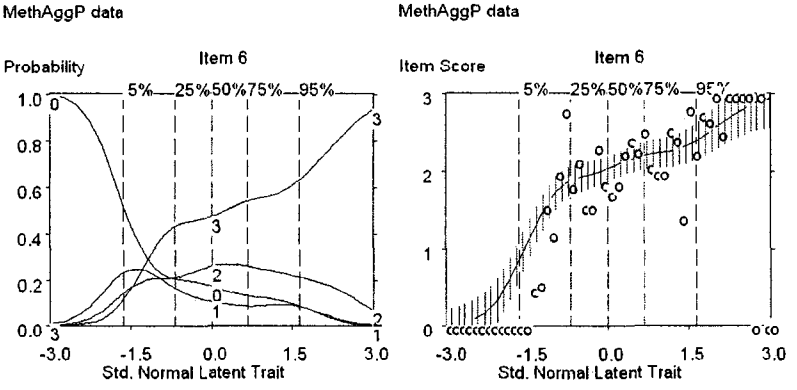


Figure 6. Option characteristic curves and item characteristic curves for the 6 items of the PAI subscale AGG-P when administered to a sample of MMT patients ($N = 323$).

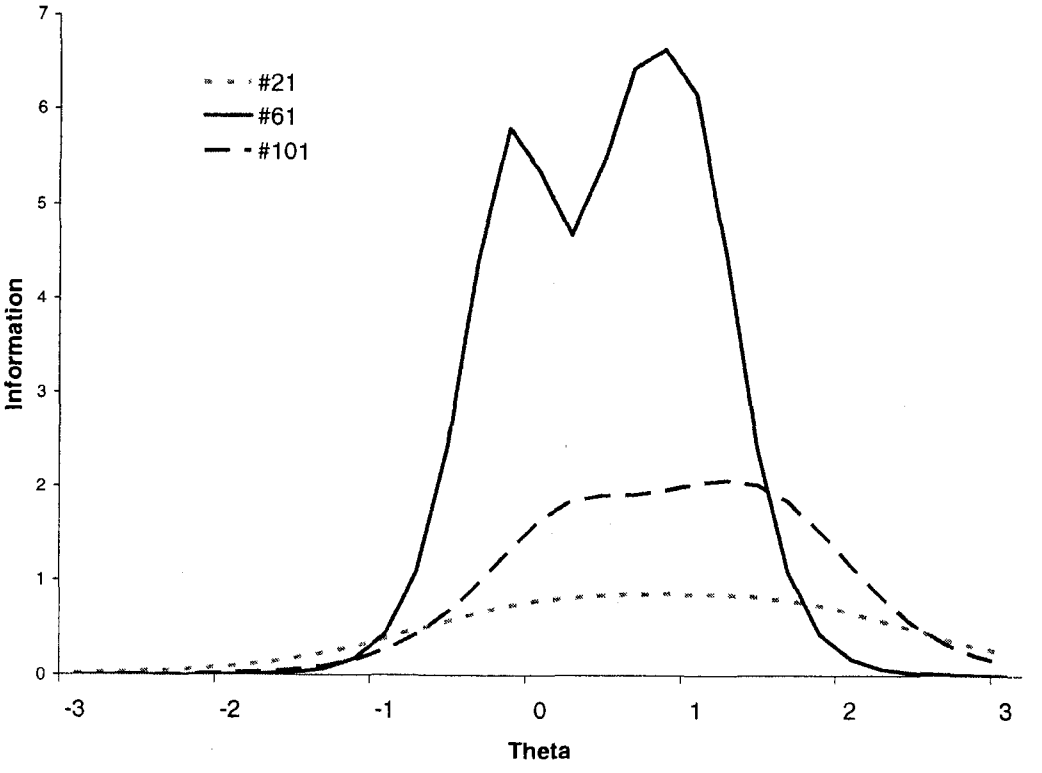


Figure 7. Comparison of item information curves for the three items of the AGG-P-R subscale when employed in a sample of MMT patients ($N = 323$).

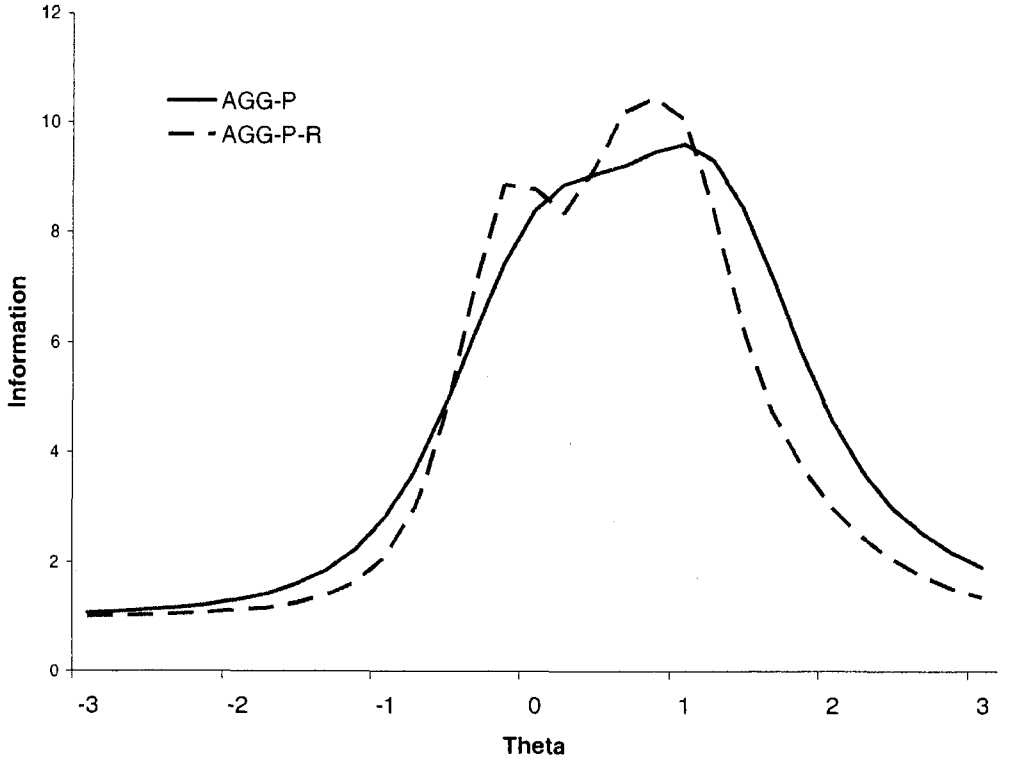
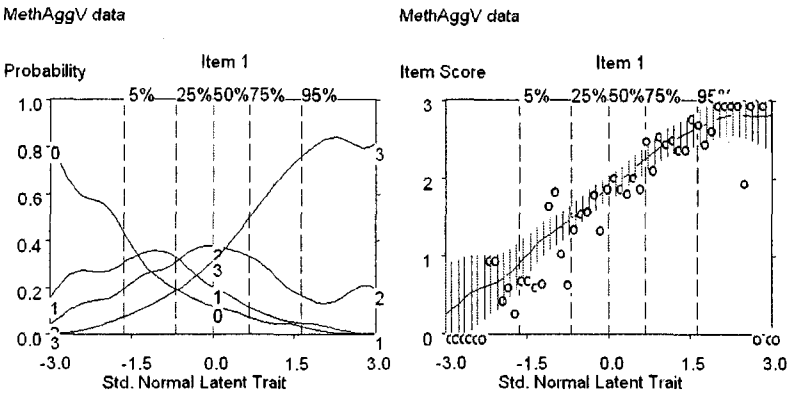
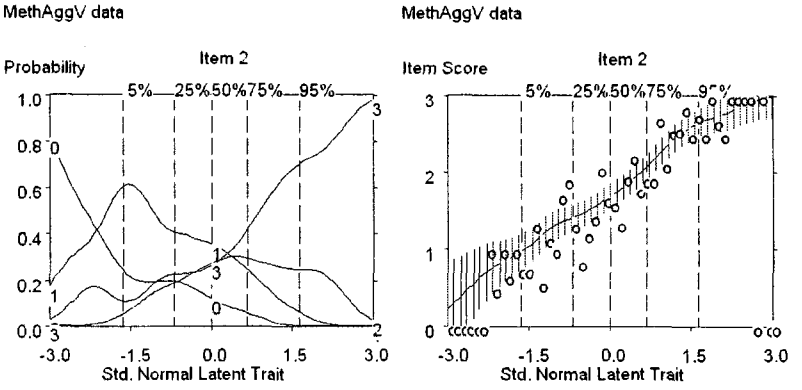


Figure 8. Comparison of the total information provided by the original 6 item AGG-P subscale and the three item AGG-P-R subscale when employed in a MMT patient population ($N = 323$).

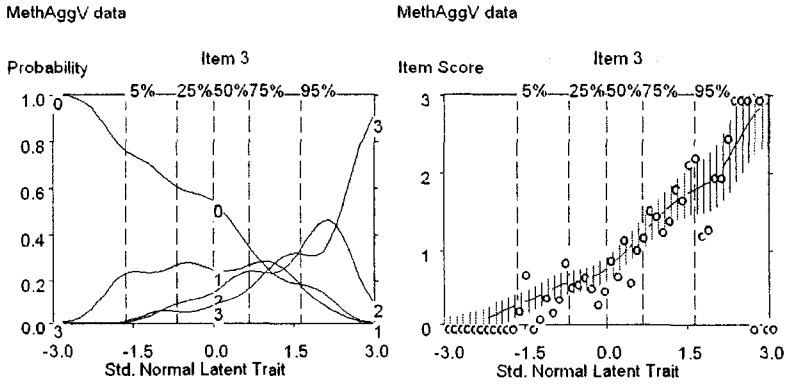
A. PAI Item #18



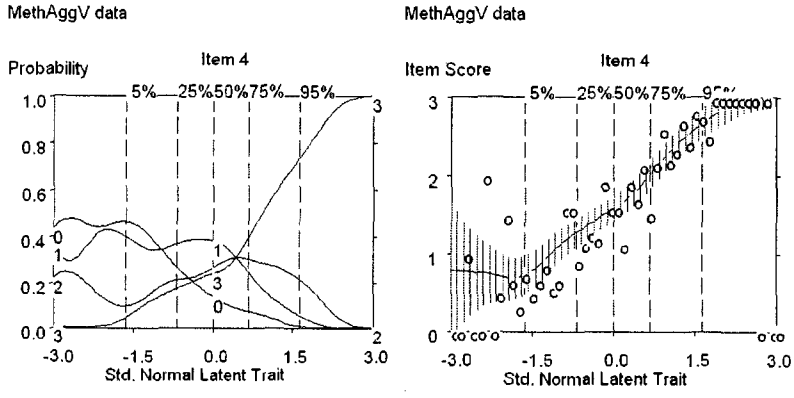
B. PAI Item #58



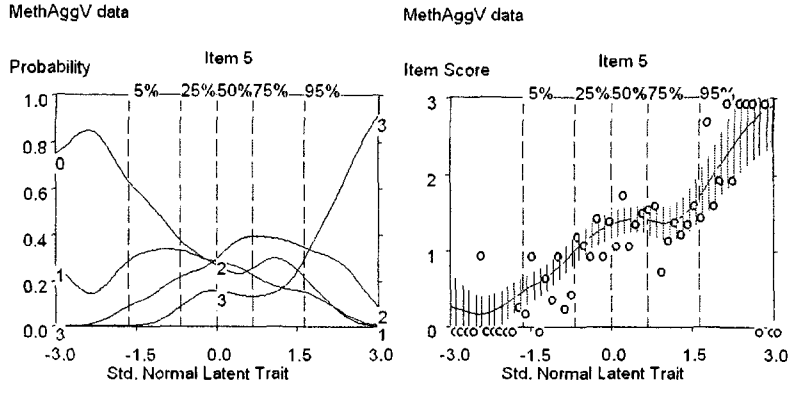
C. PAI Item #98



D. PAI Item #138



E. PAI Item #178



PAI Item #218

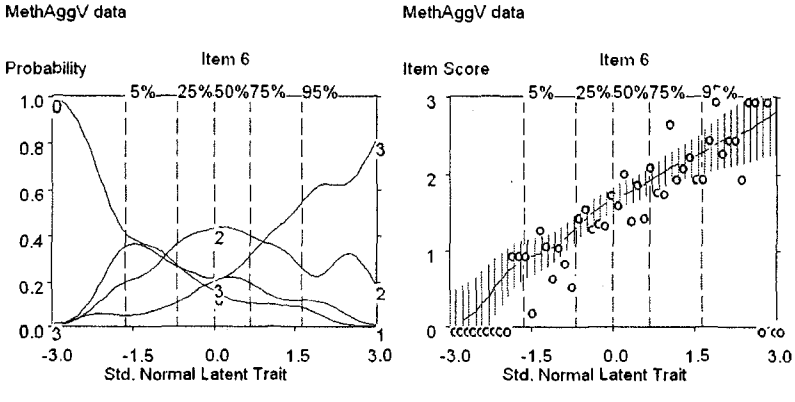


Figure 9. Option and item characteristic curves for the PAI subscale AGG-V based upon responses from a sample of MMT patients ($N = 323$).

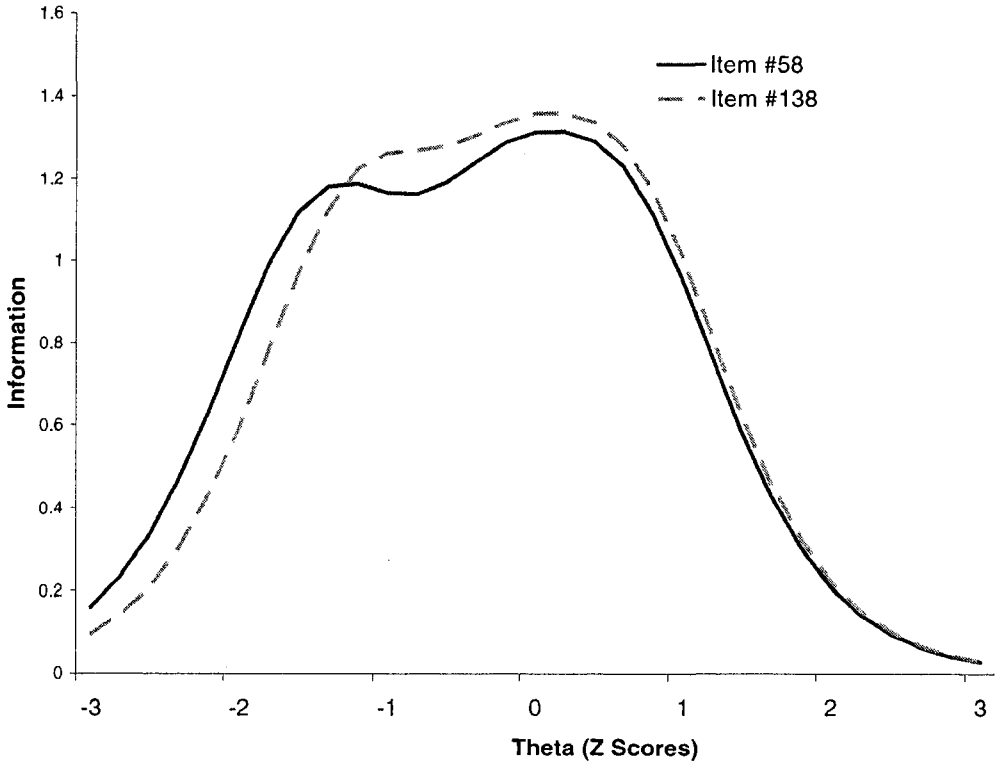


Figure 10. Item information curves for the two items of the AGG-V-R subscale when used in a MMT patient population ($N = 323$).

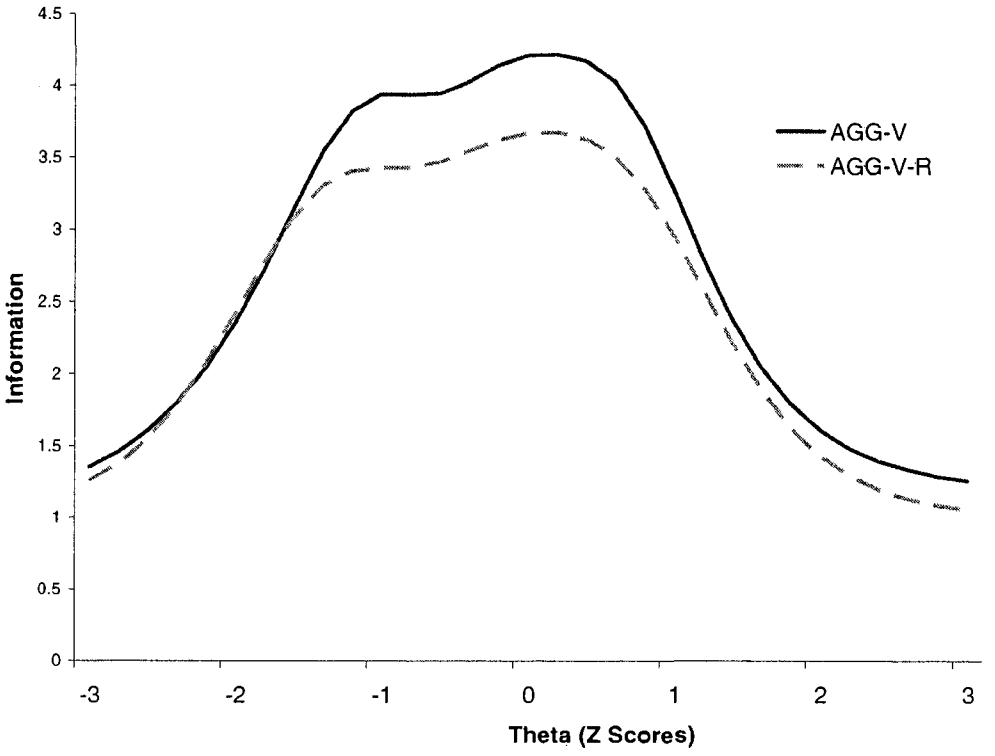
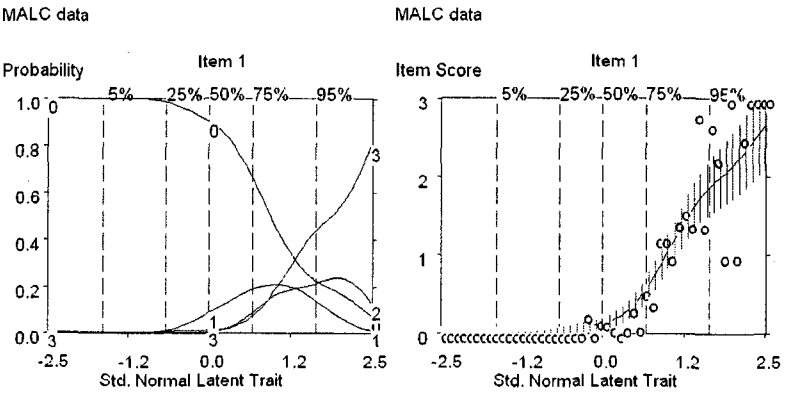
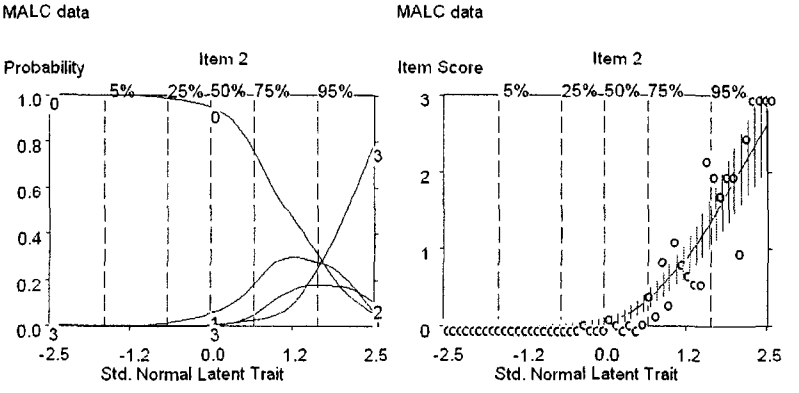


Figure 11. Comparison of total information of the original AGG-V (6 Items) to the AGG-V-R (2 Items) in a MMT sample ($N = 323$).

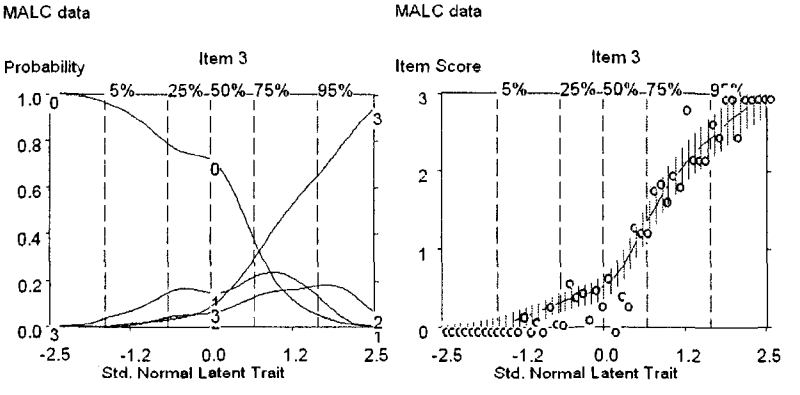
PAI Item #15



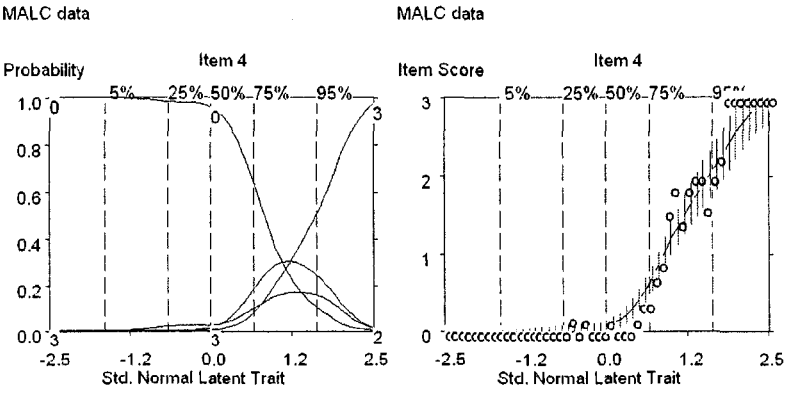
PAI Item #55



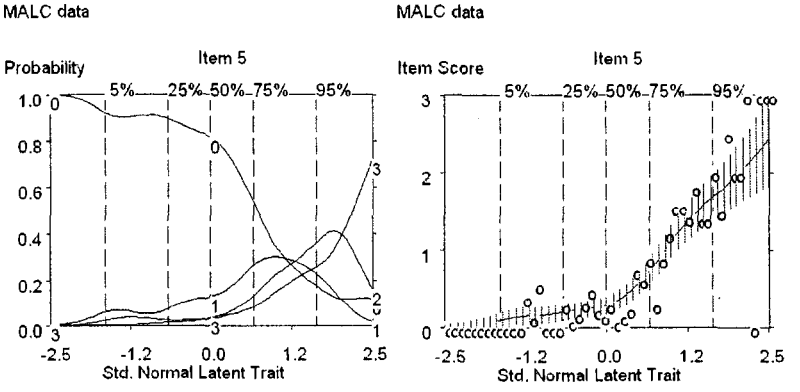
PAI Item #95



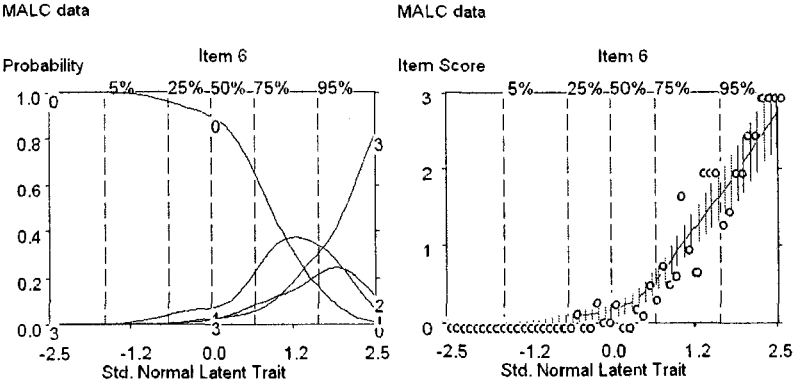
PAI Item #135



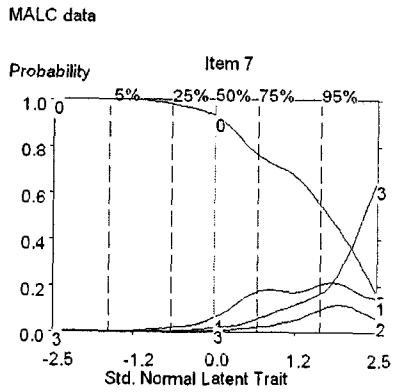
PAI Item #175

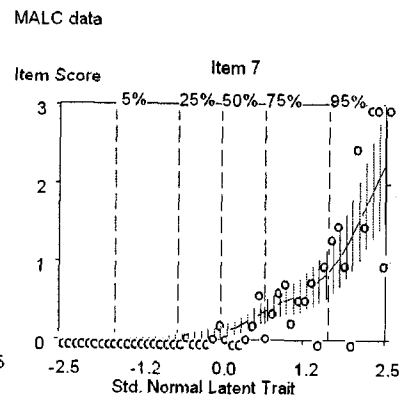


PAI Item #215

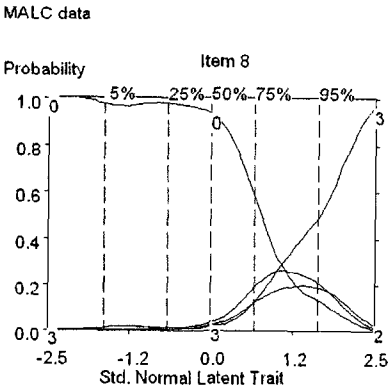


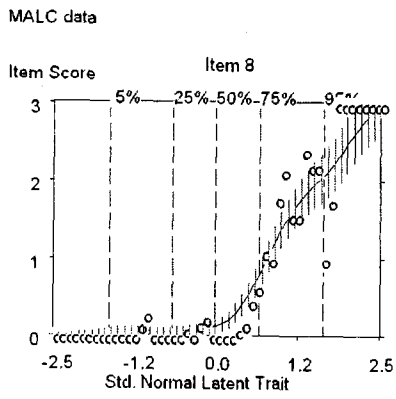
PAI Item #254



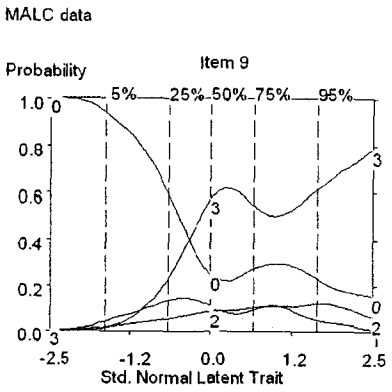


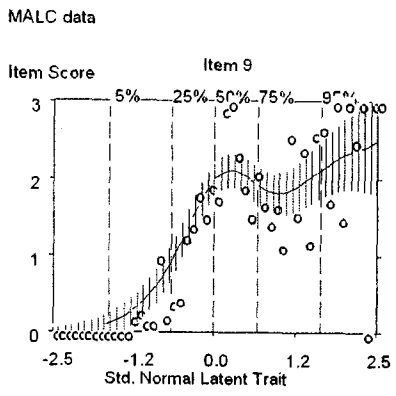
PAI Item #255



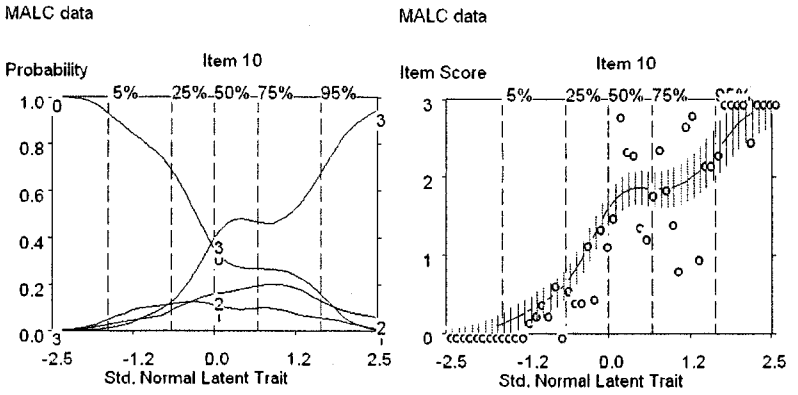


PAI Item #294

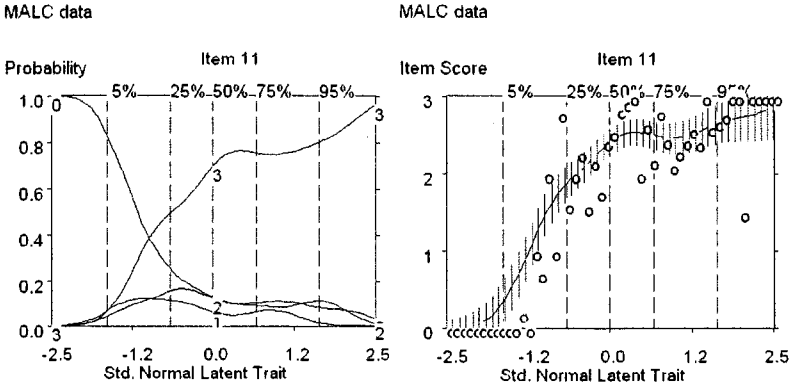




PAI Item #295



PAI Item #334



PAI Item #335

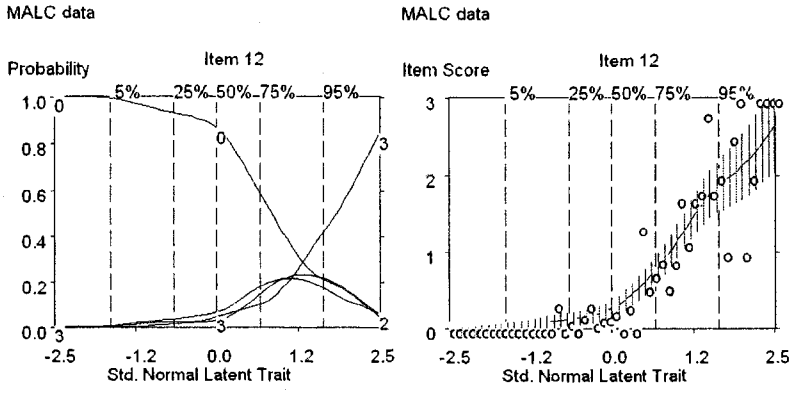


Figure 12. Item and option characteristic curves are display for the ALC scale in a sample of MMT patients (N = 323).

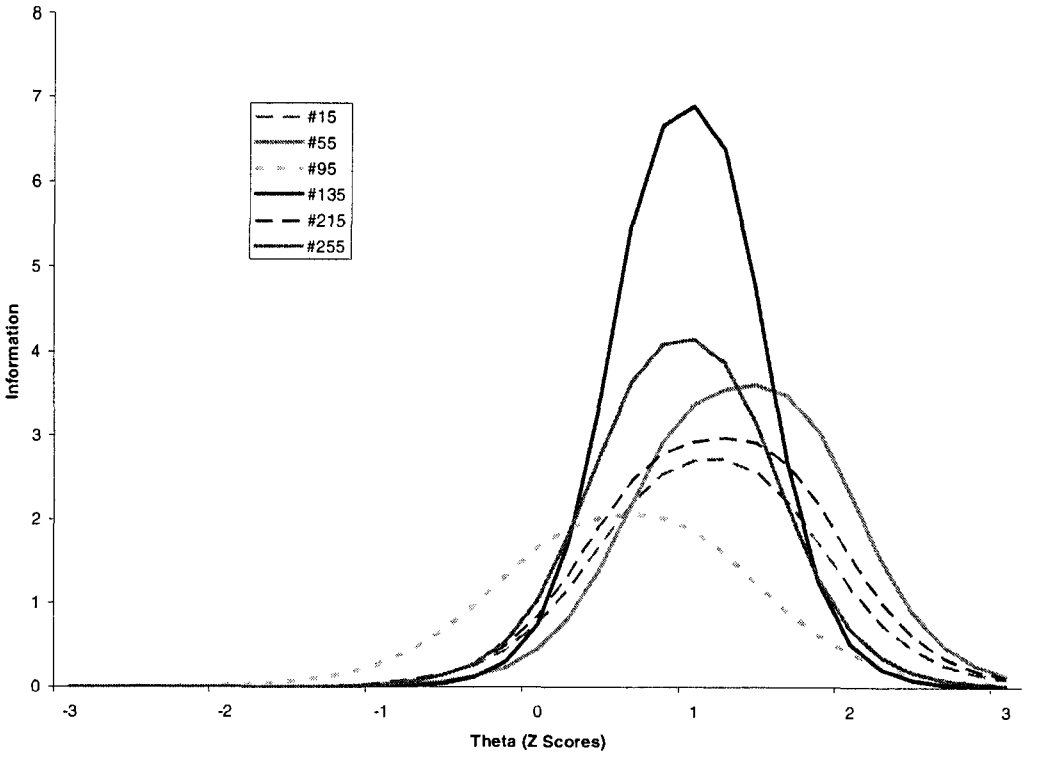


Figure 13. Individual item information curves for the ALC-R scale in a sample of MMT patients ($N = 323$).

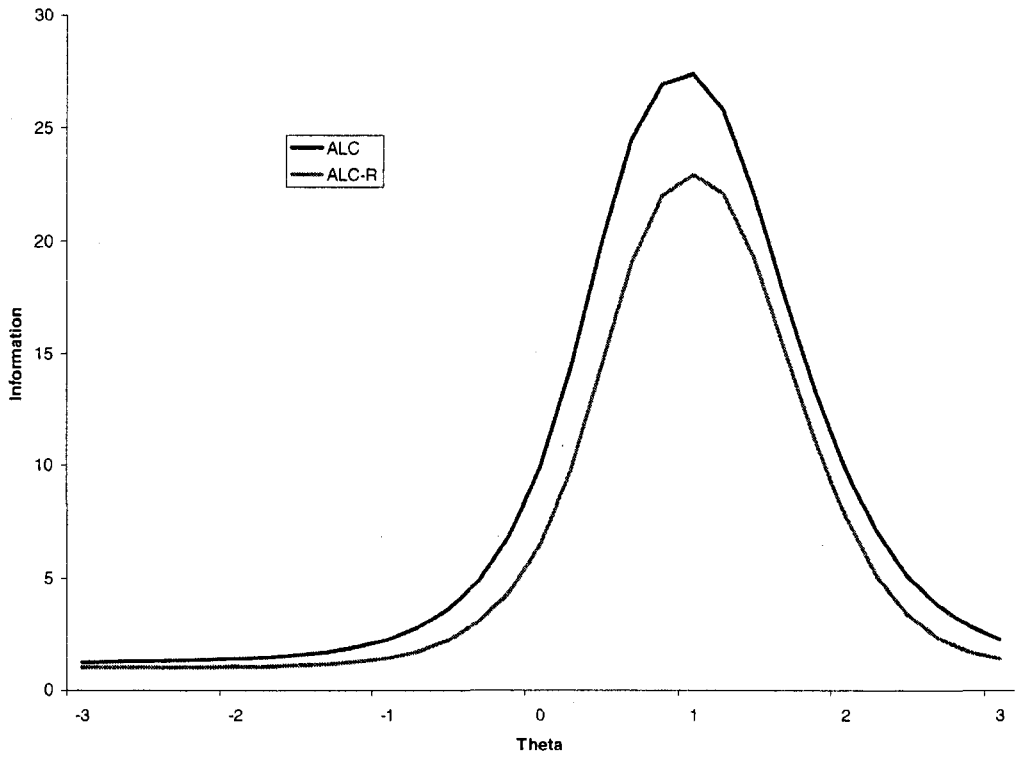
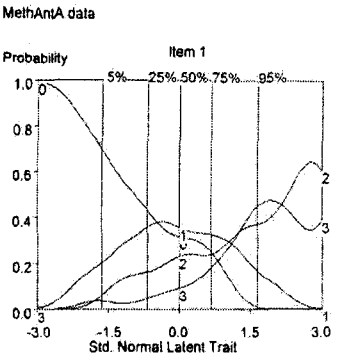
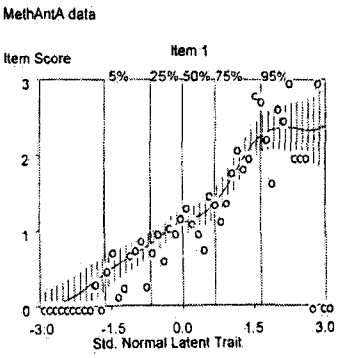


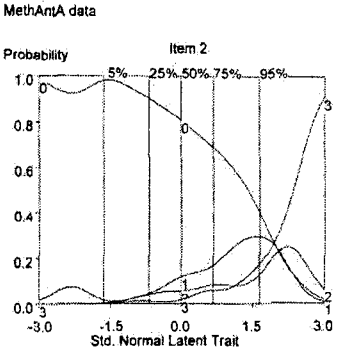
Figure 14. Total scale information curves comparing the original 12 item ALC scale to the 6 item ALC-R scale in a sample of MMT patients (N = 323).

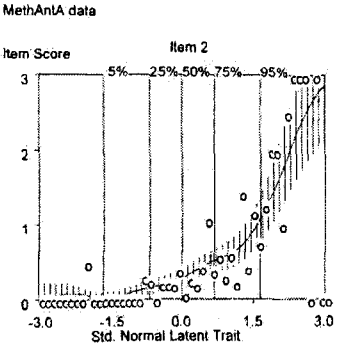
PAI Item #11



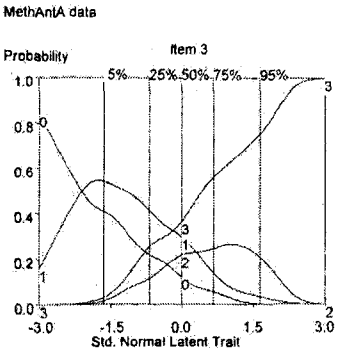


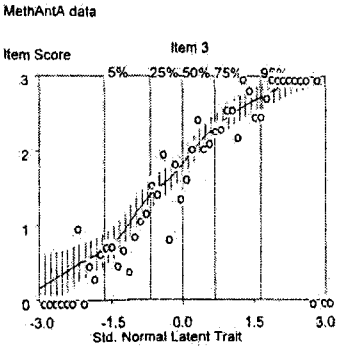
PAI Item # 51



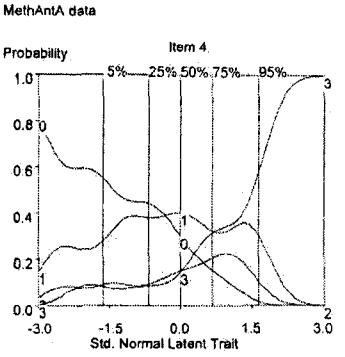


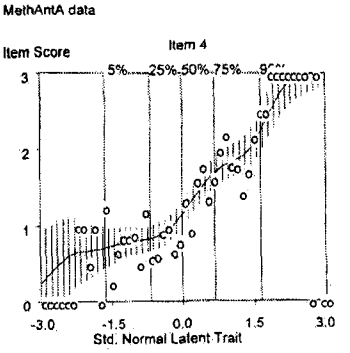
PAI Item #91



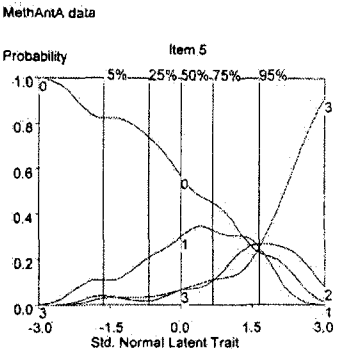


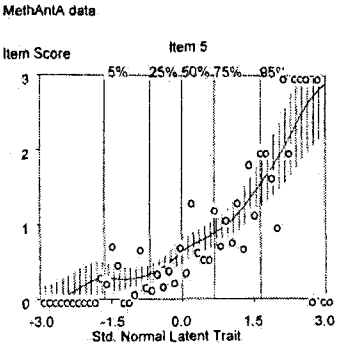
PAI Item # 141



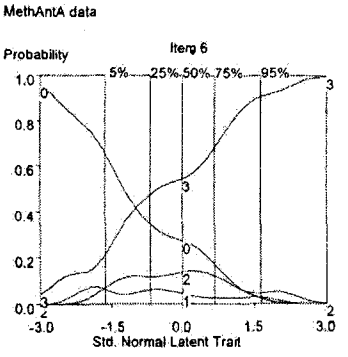


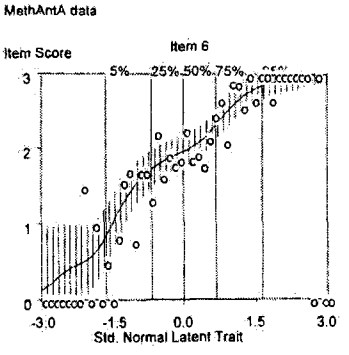
PAI Item # 171



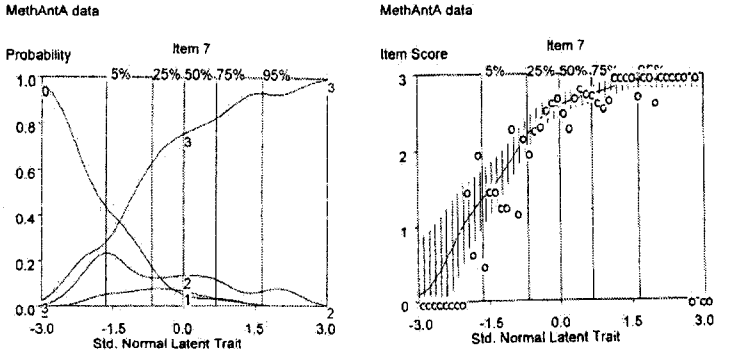


PAI Item #211





PAI Item #251



PAI Item #291

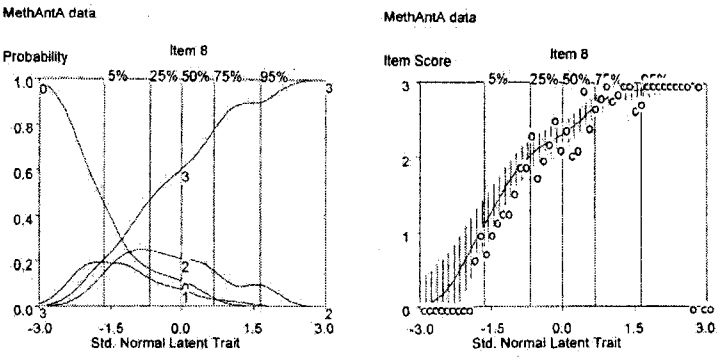


Figure 15. Option and item characteristics were computed for the items of the ANT-A subscale when employed in a sample of MMT patients ($N = 323$).

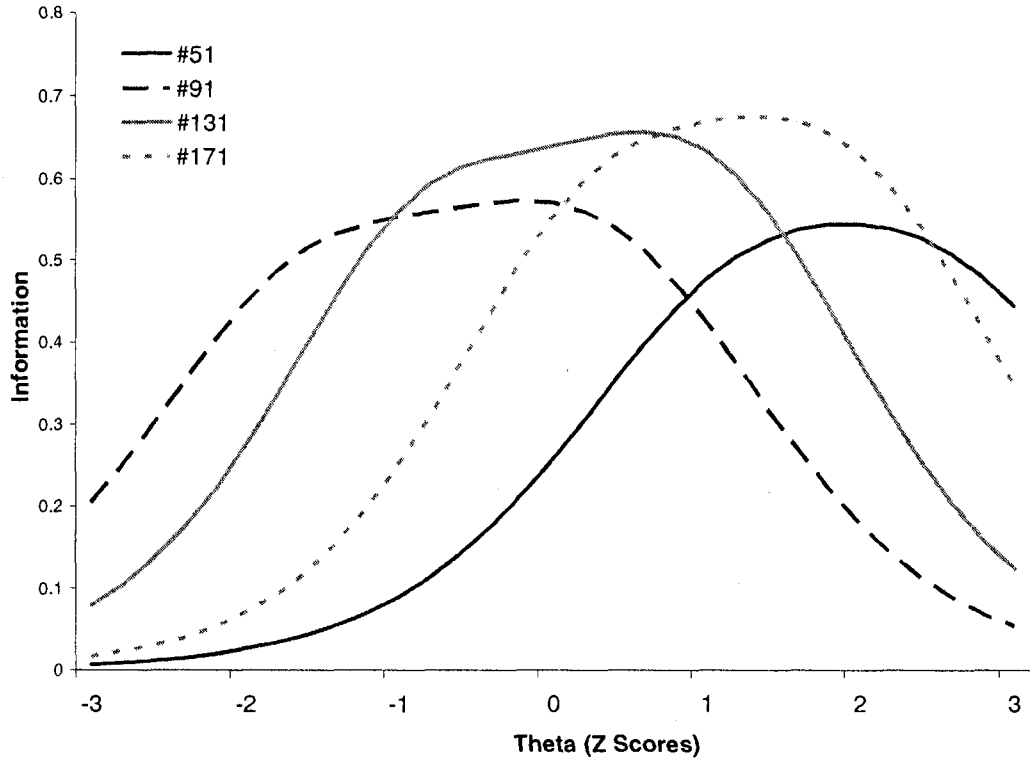


Figure 16. Item information curves for the ANT-A-R subscale when administered to a sample of MMT patients ($N = 323$).

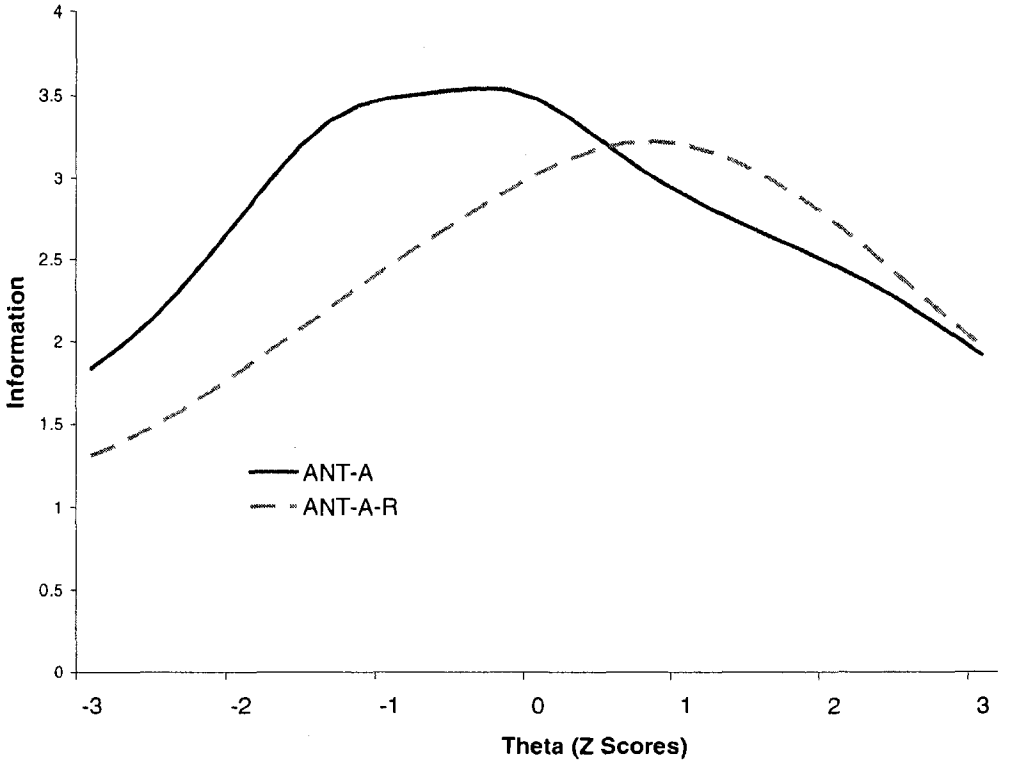
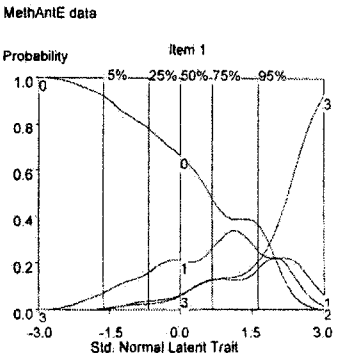
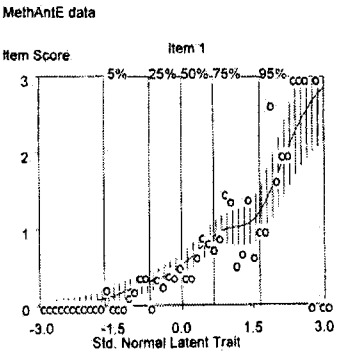


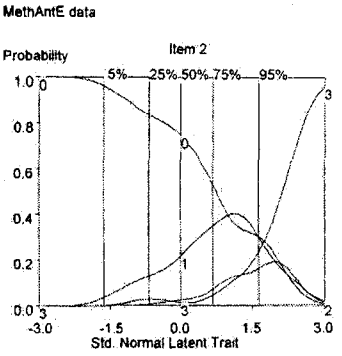
Figure 17. Comparison of the information curves of the original ANT-A subscale (8 items) and the ANT-A-R subscale when responded to by a sample of MMT patients (N = 323).

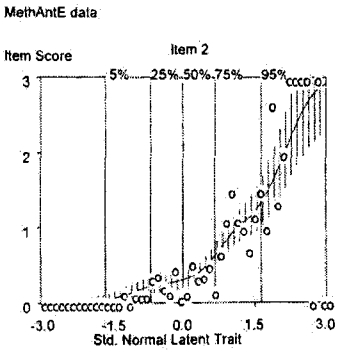
PAI Item # 31



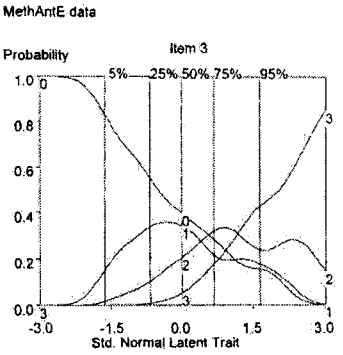


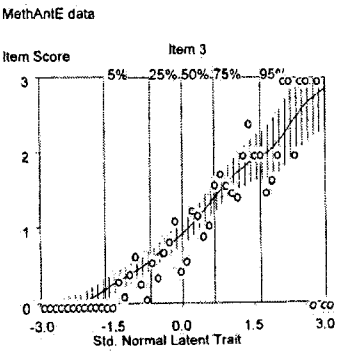
PAI Item # 71



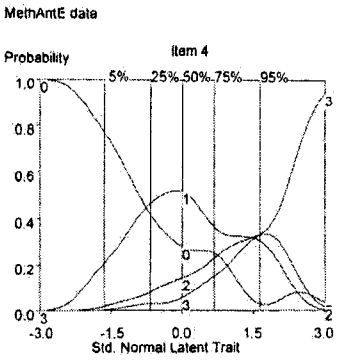


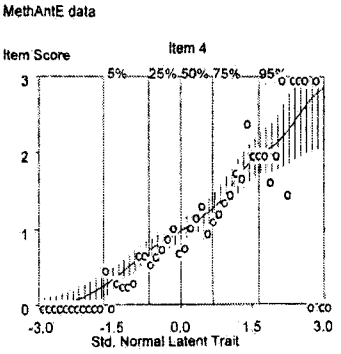
PAI Item # 111



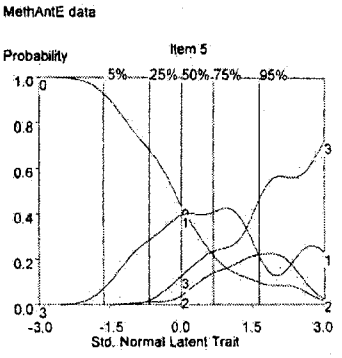


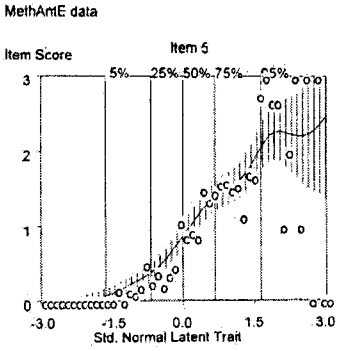
PAI Item # 151



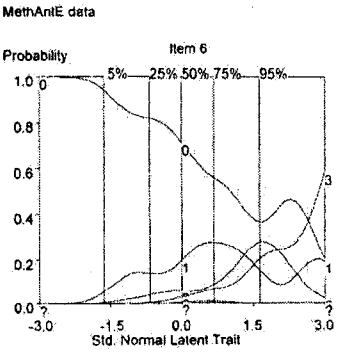


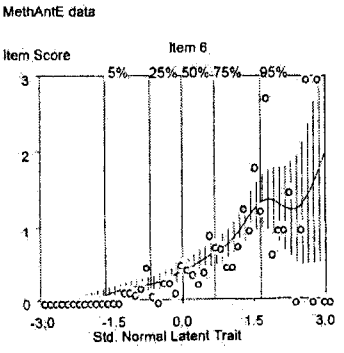
PAI Item # 191



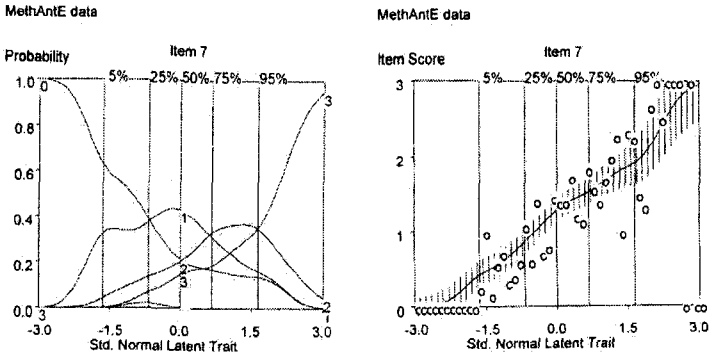


PAI Item # 231





PAI Item # 271



PAI Item #311

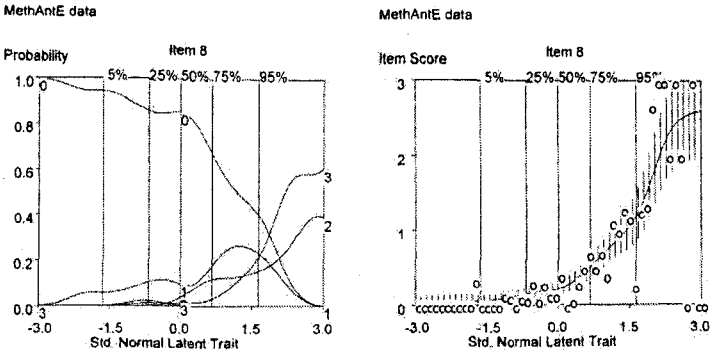


Figure 18. Item and option characteristic curves for the ANT-E subscale of the PAI as calibrated by the NIRT Testgraf software when employed in a MMT patient population (N = 323).

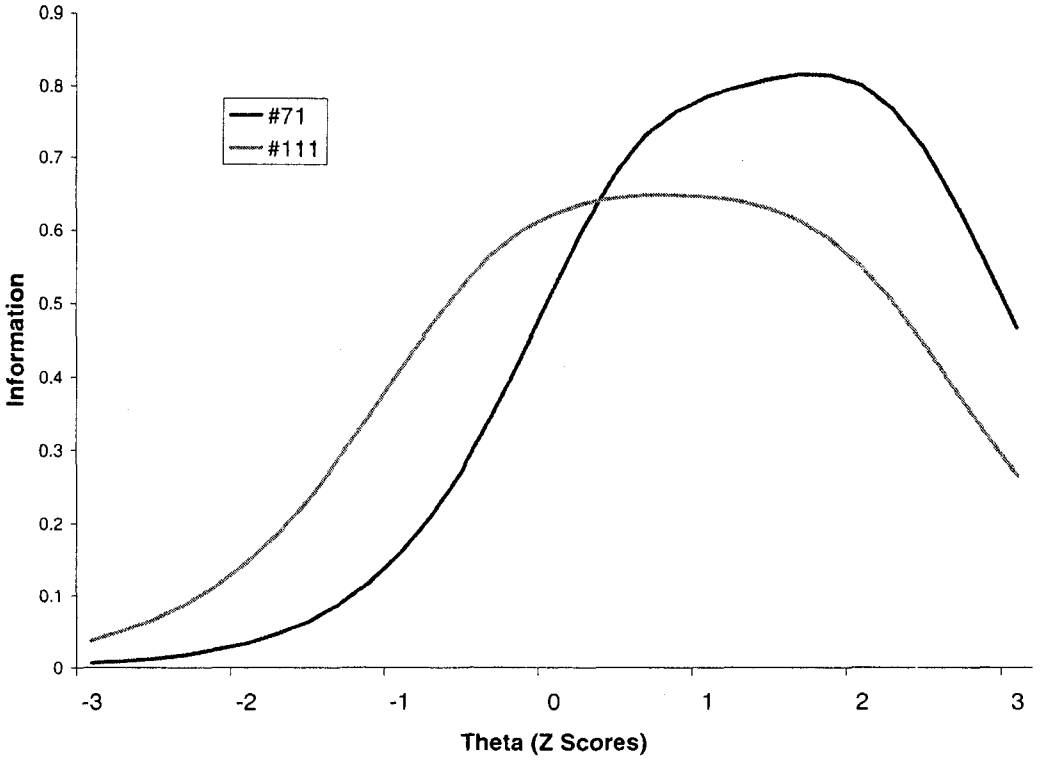
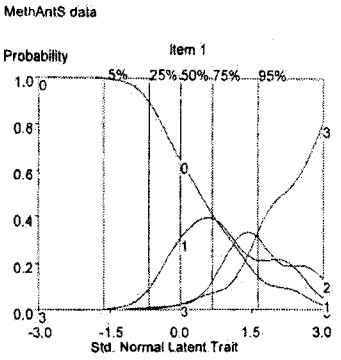
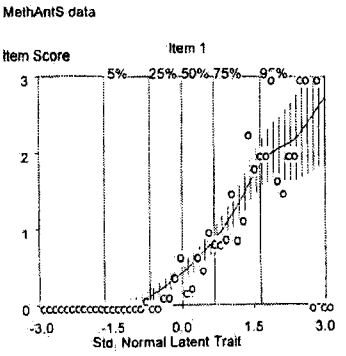


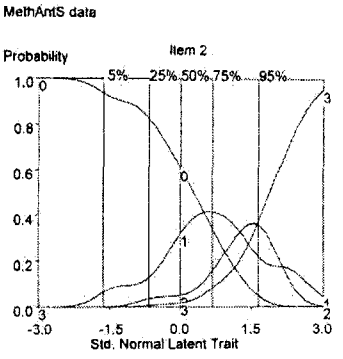
Figure 19. Item information curves for the two remaining items of the ANT-E-R based upon responses to the PAI by a sample of MMT patients ($N = 323$).

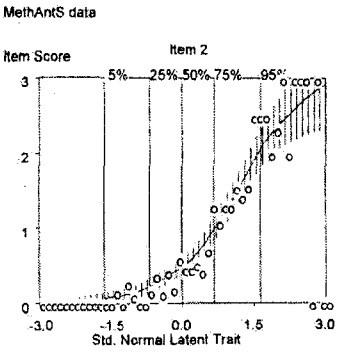
PAI Item # 39



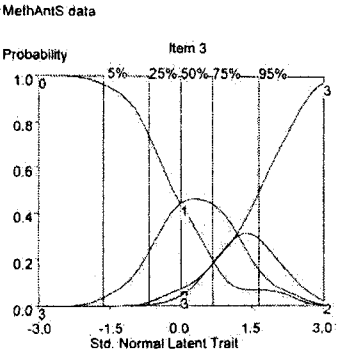


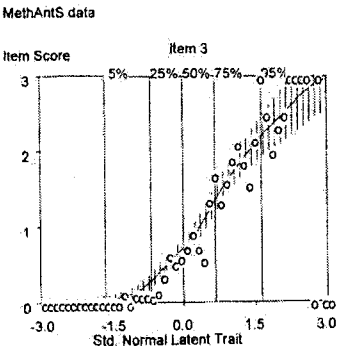
PAI Item # 79



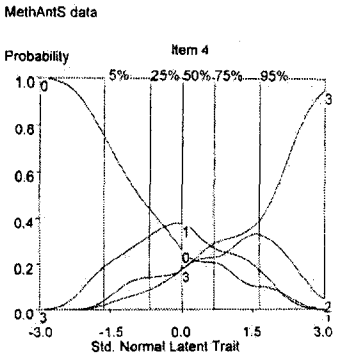


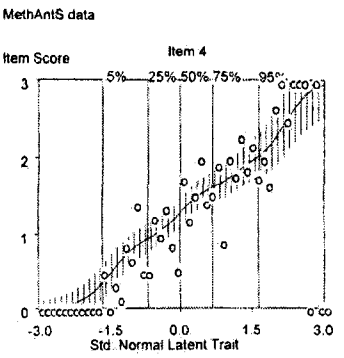
PAI Item # 119



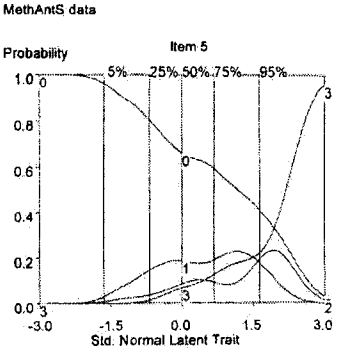


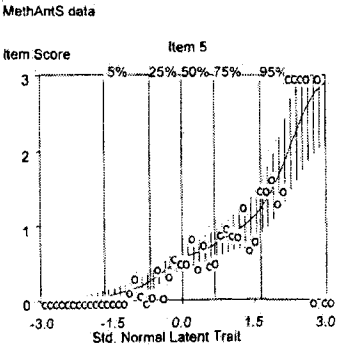
PAI Item # 159



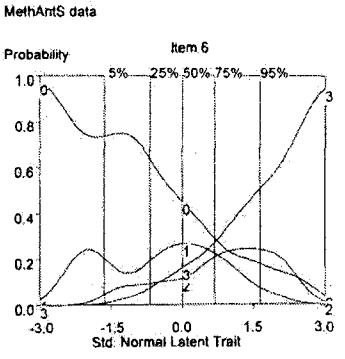


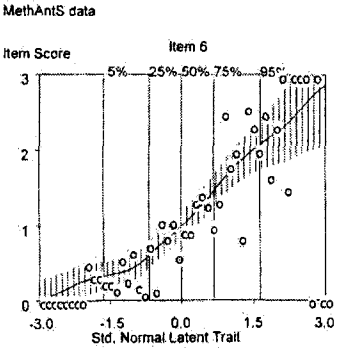
PAI Item # 199





PAI Item # 239





PAI Item # 279

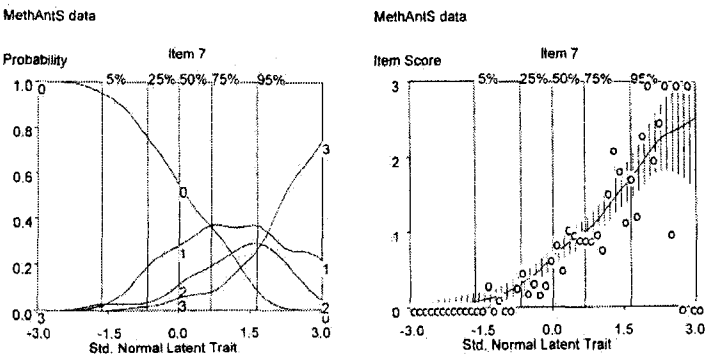


Figure 20. Item and option characteristic curves for the ANT-S subscale of the PAI when used in a sample of MMT patients ($N = 323$).

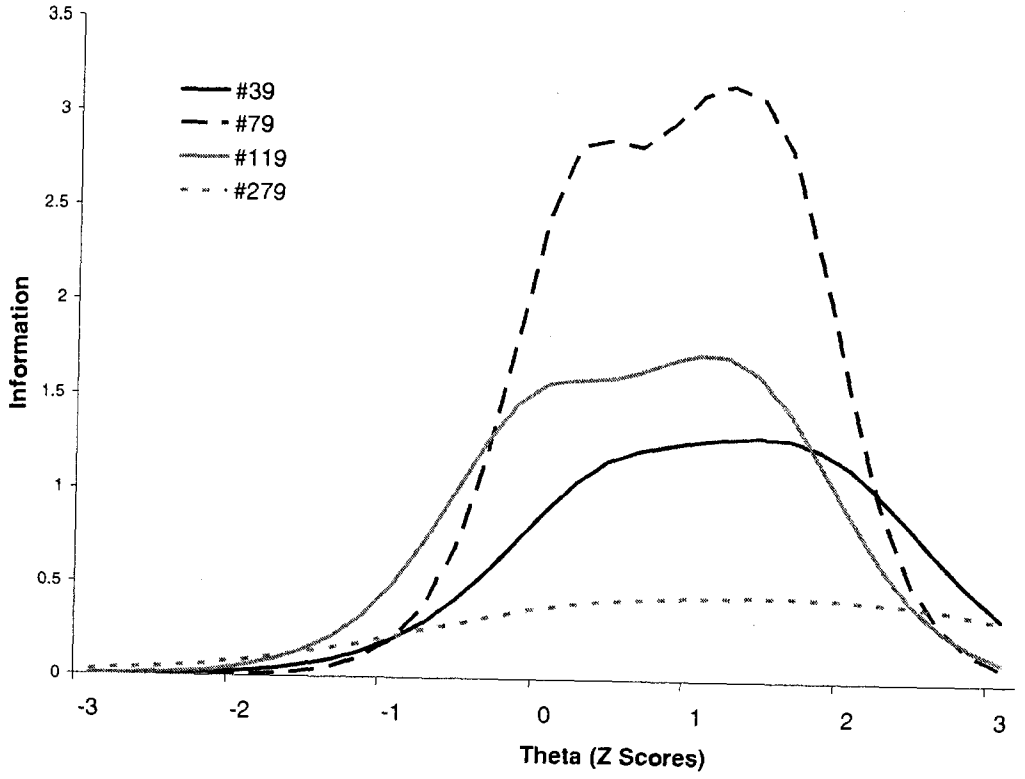


Figure 21. Item information curves for the ANT-S-R subscale when used in a sample of MMT patients ($N = 323$).

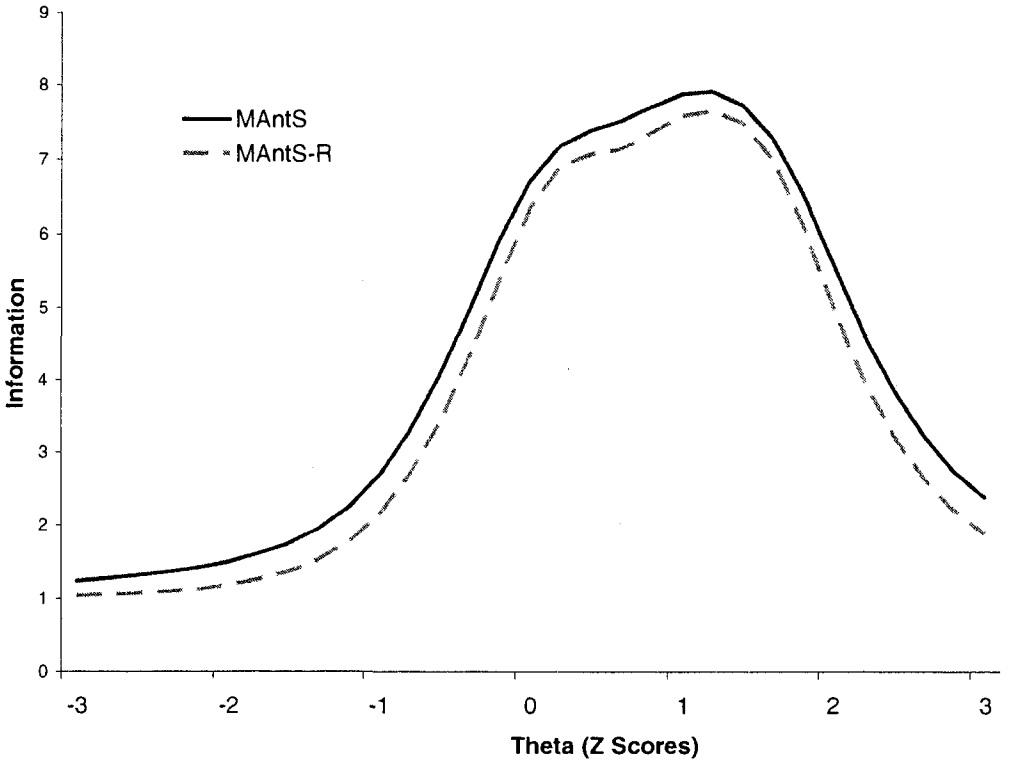
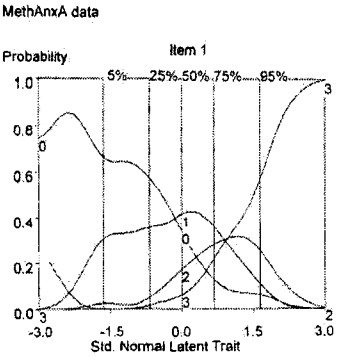
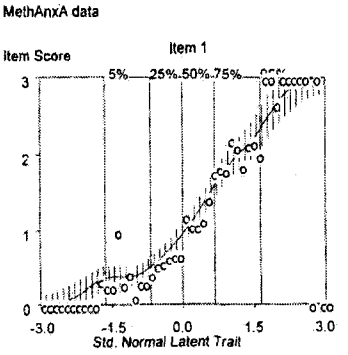


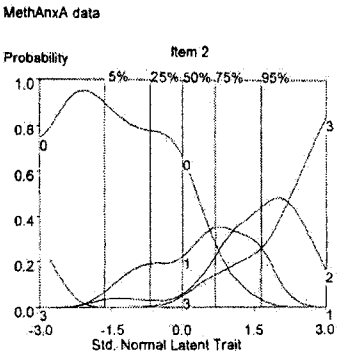
Figure 22. A comparison of the total information provided by the 8 item ANT-S subscale compared to the 4 item ANT-S-R subscale when employed in a MMT sample ($N = 323$).

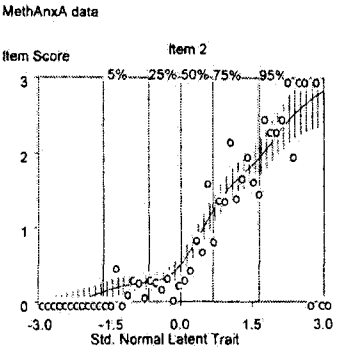
PAI Item #4



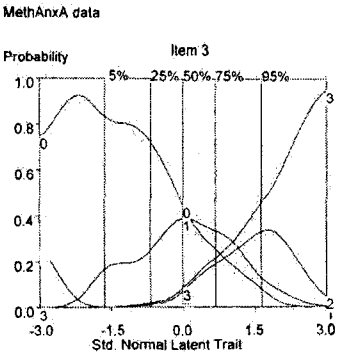


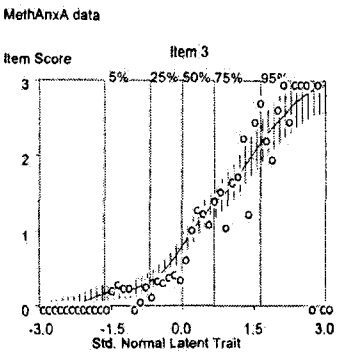
PAI Item #44



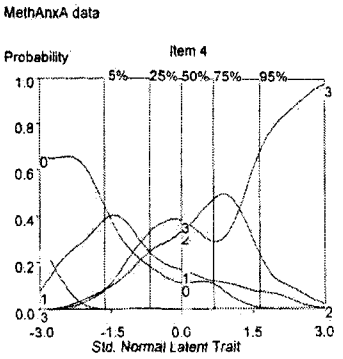


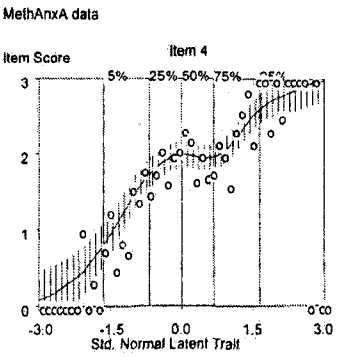
PAI Item #84



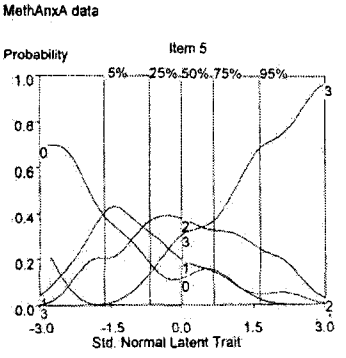


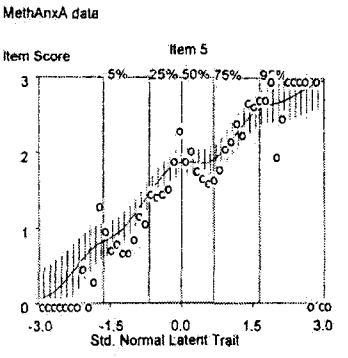
PAI Item # 124



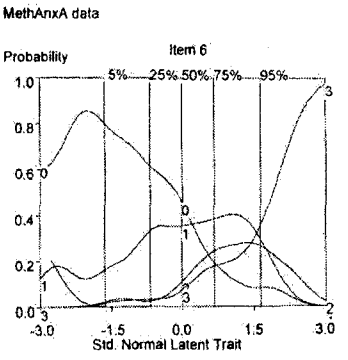


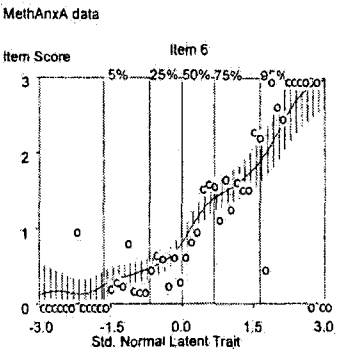
PAI Item #164



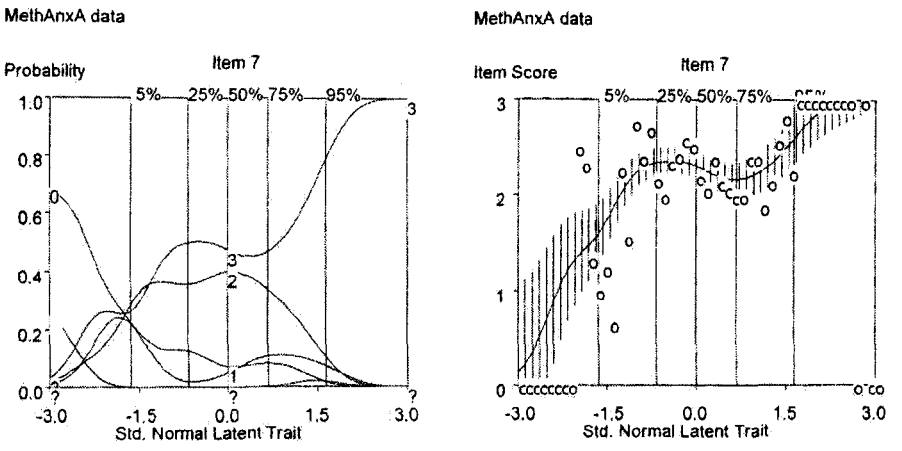


PAI Item # 204





PAI Item #244



PAI Item #284

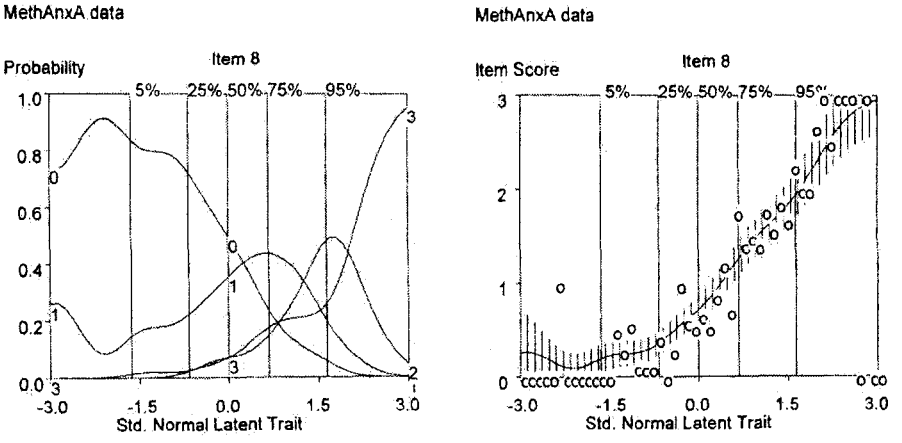


Figure 23. Option and item characteristic curves for the ANX-A subscale as computed by Testgraf based upon responses by a MMT sample (N = 323).

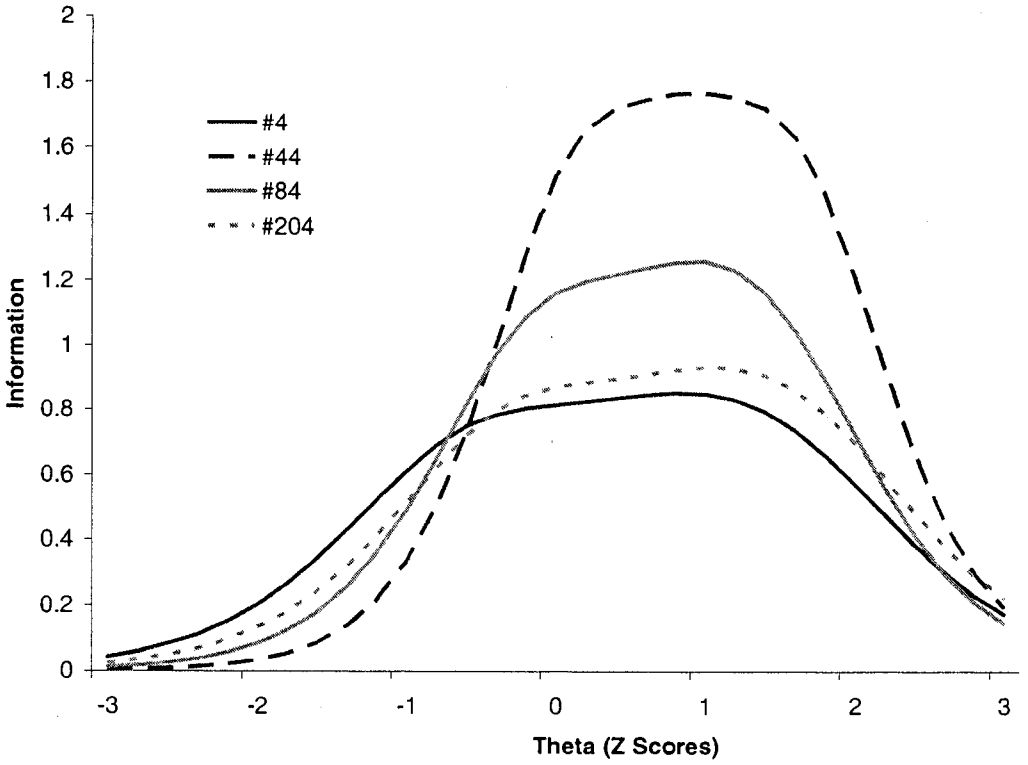


Figure 24. Item information curves for the four items of the ANX-A-R subscale in a MMT patient population ($N = 323$).

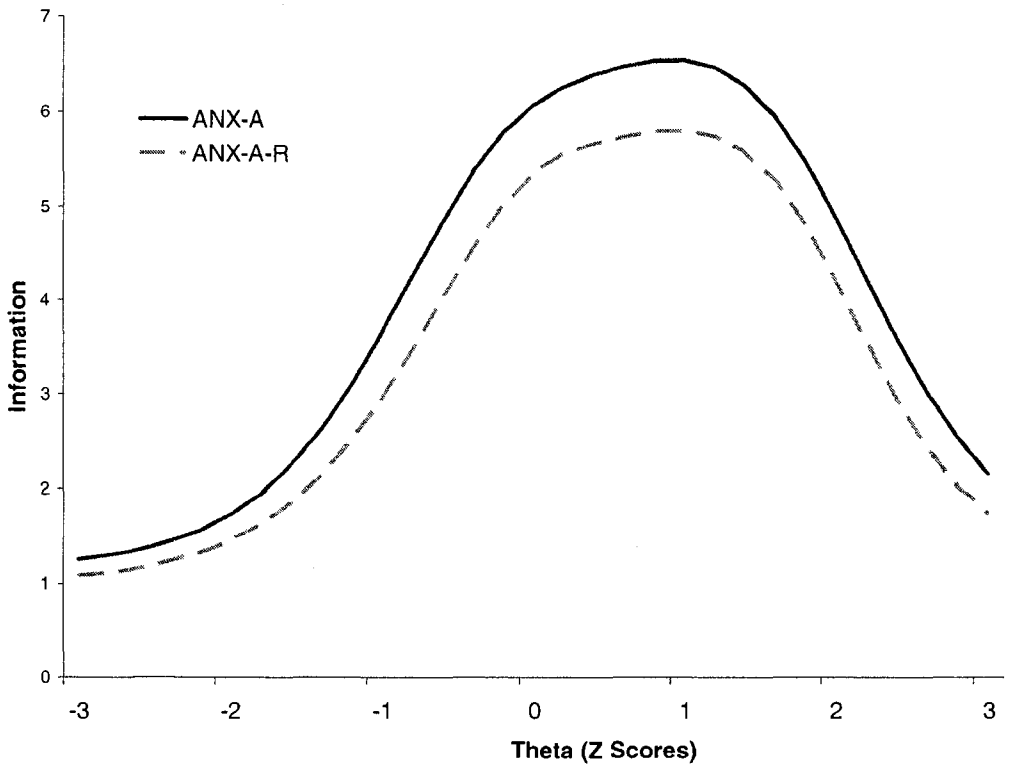
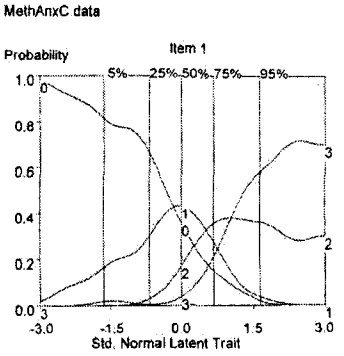
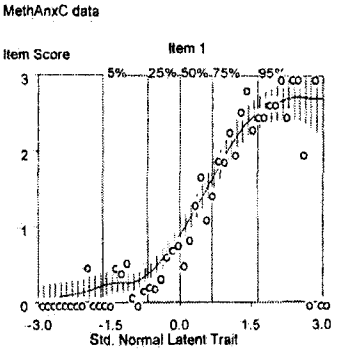


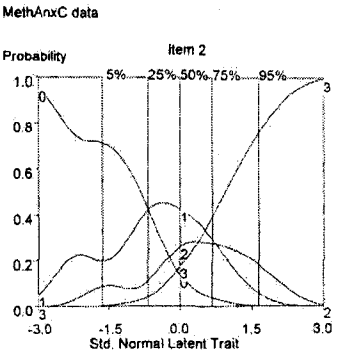
Figure 25. Comparison of total information for the eight item ANX-A subscale and the four item ANX-A-R subscale based upon responses from a MMT patient sample (N = 323).

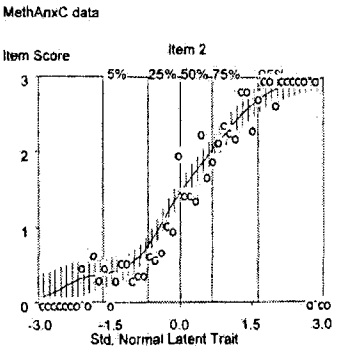
PAI Item #25



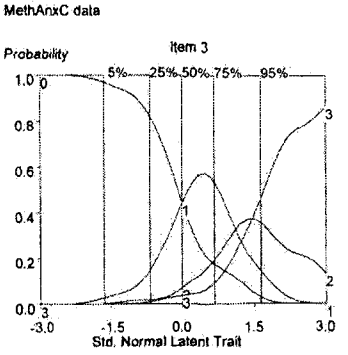


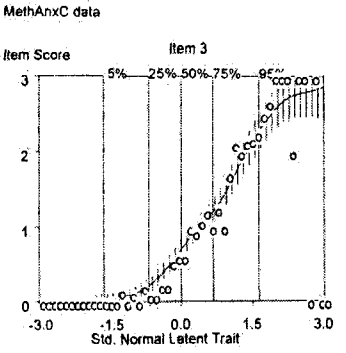
PAI Item #65



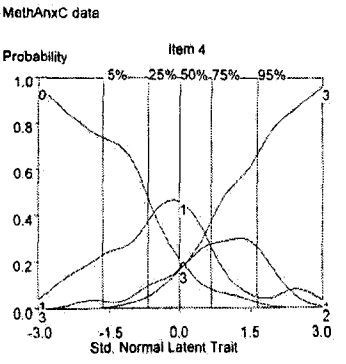


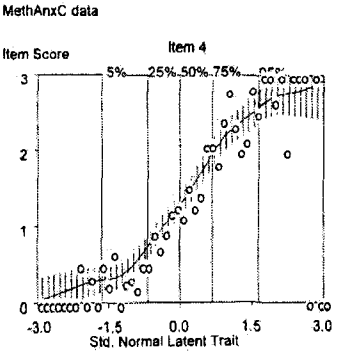
PAI Item #105



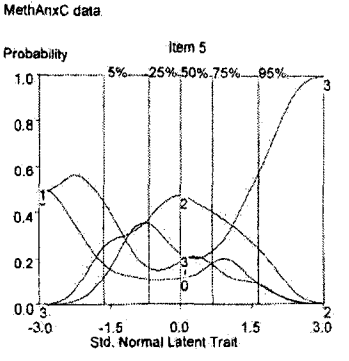


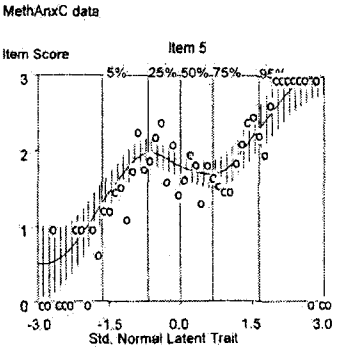
PAI Item #145



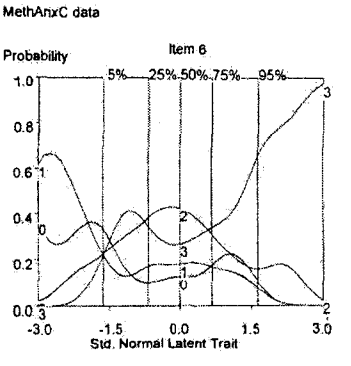


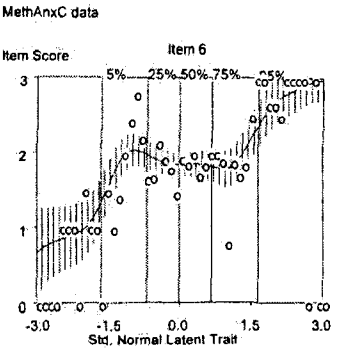
PAI Item #185



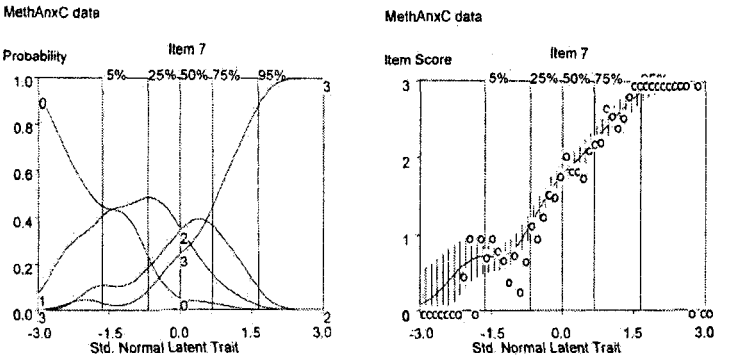


PAI Item #225





PAI Item #265



PAI Item #305

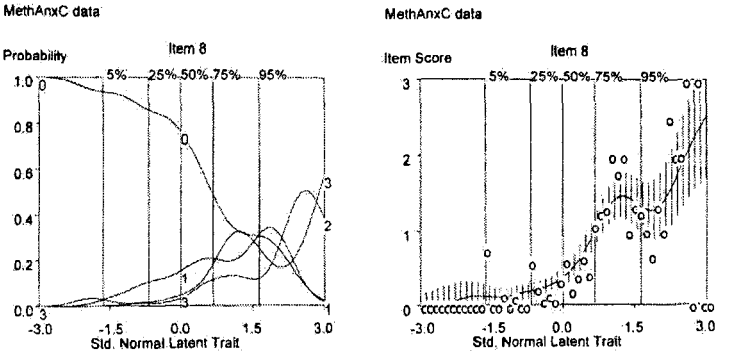


Figure 26. Option and information curves for the ANX-C subscale of the PAI based upon responses from a sample of MMT patients (N = 323).

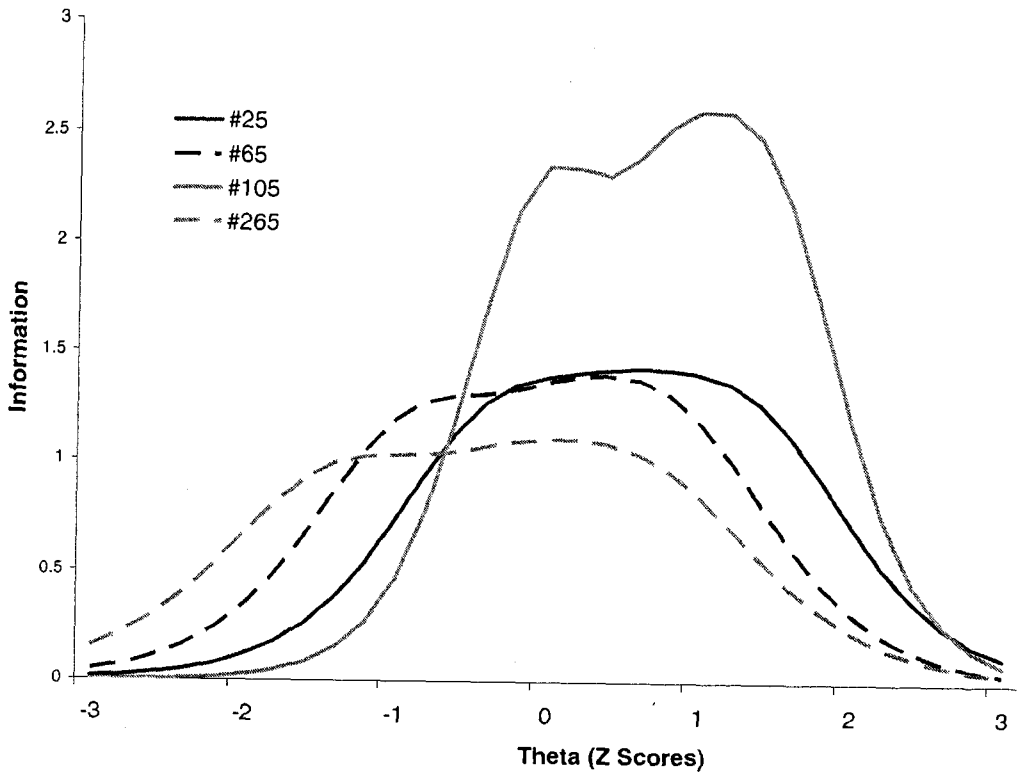


Figure 27. Item information curves for the new ANX-C-R subscale when employed in an MMT patient population ($N = 323$).

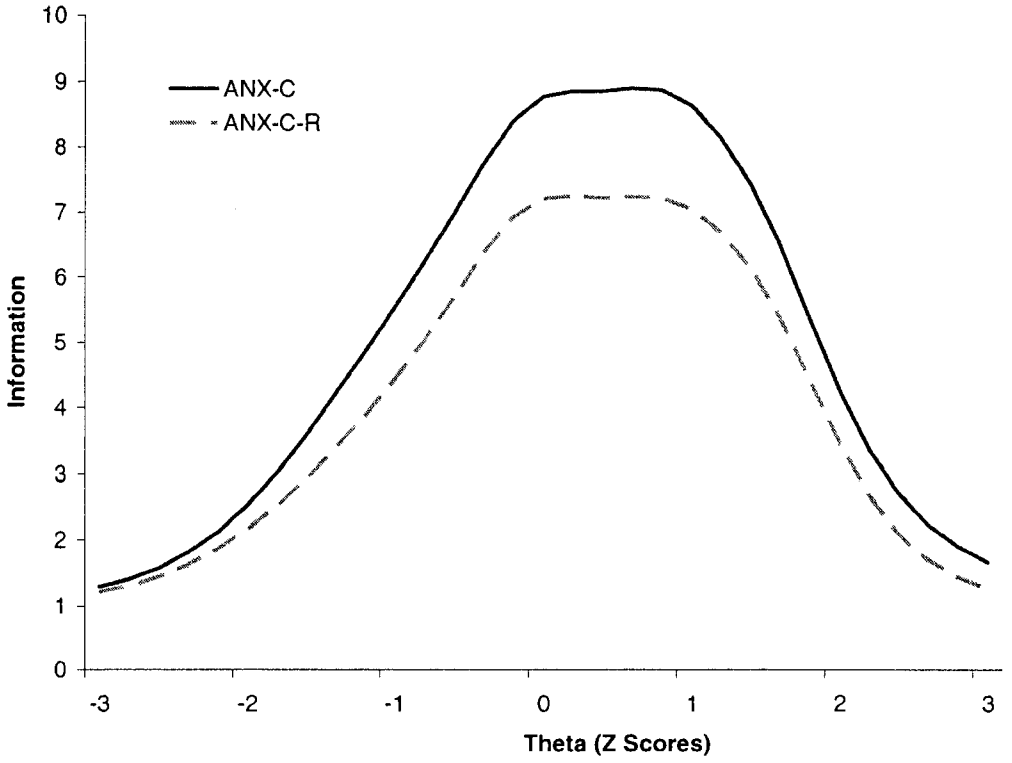
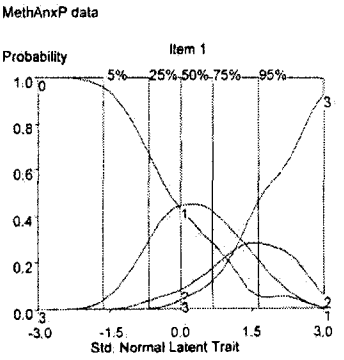
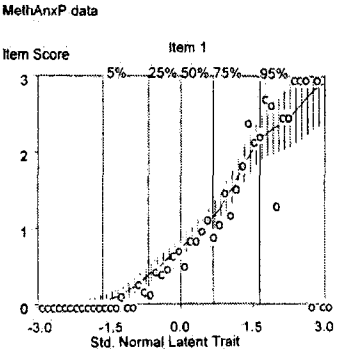


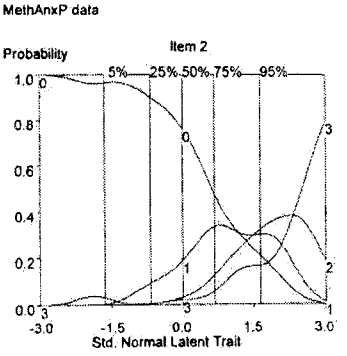
Figure 28. Comparison of total scale information of the original ANX-C subscale and the new ANX-C-R subscale when used in a sample of MMT patients ($N = 323$).

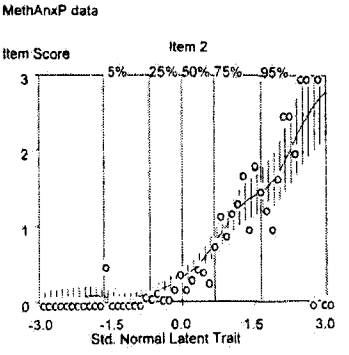
PAI Item #33



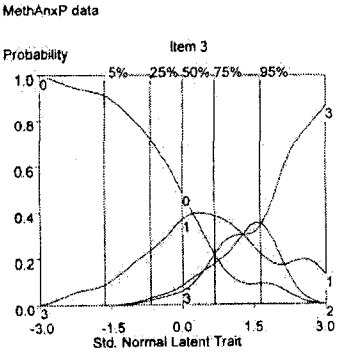


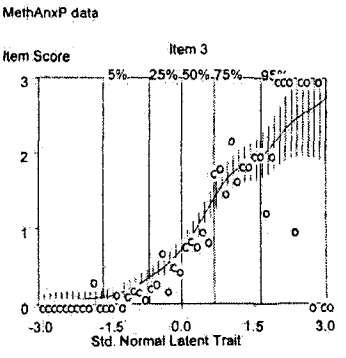
PAI Item #73



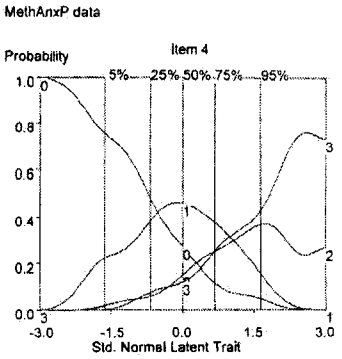


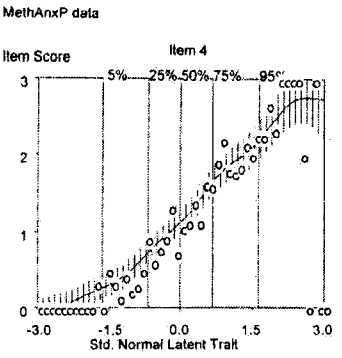
PAI Item #113



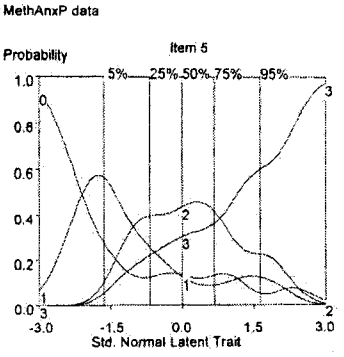


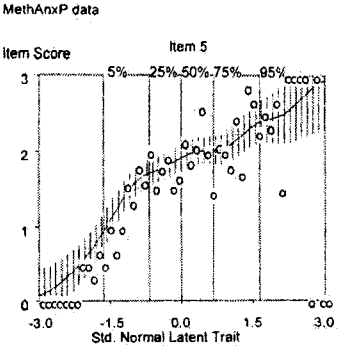
PAI Item #153



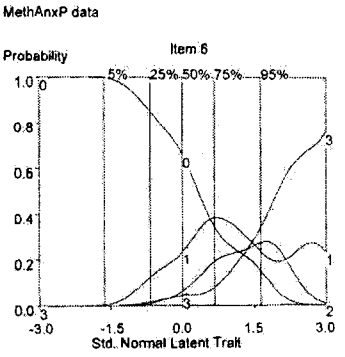


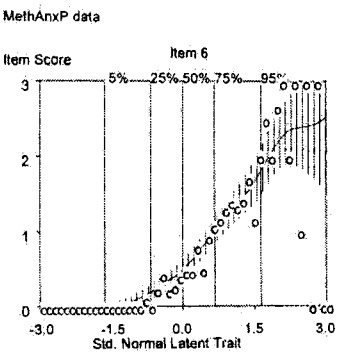
PAI Item #193



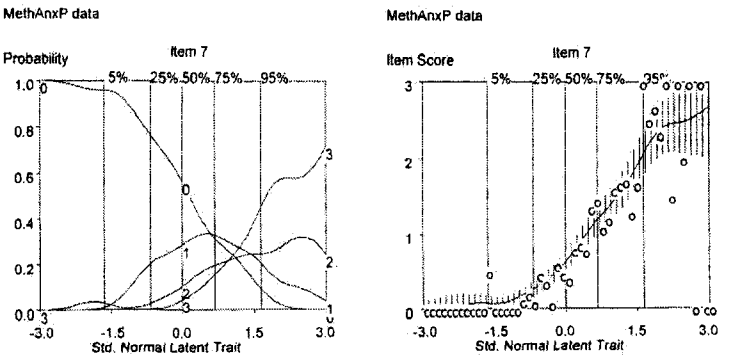


PAI Item # 233





PAI Item #273



PAI Item #313

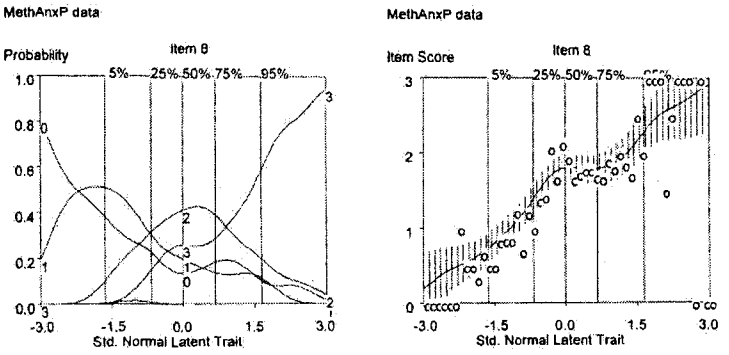


Figure 29. Option and item information curves for the ANX-P subscale of the PAI based upon responses from a sample of MMT patients ($N = 323$).

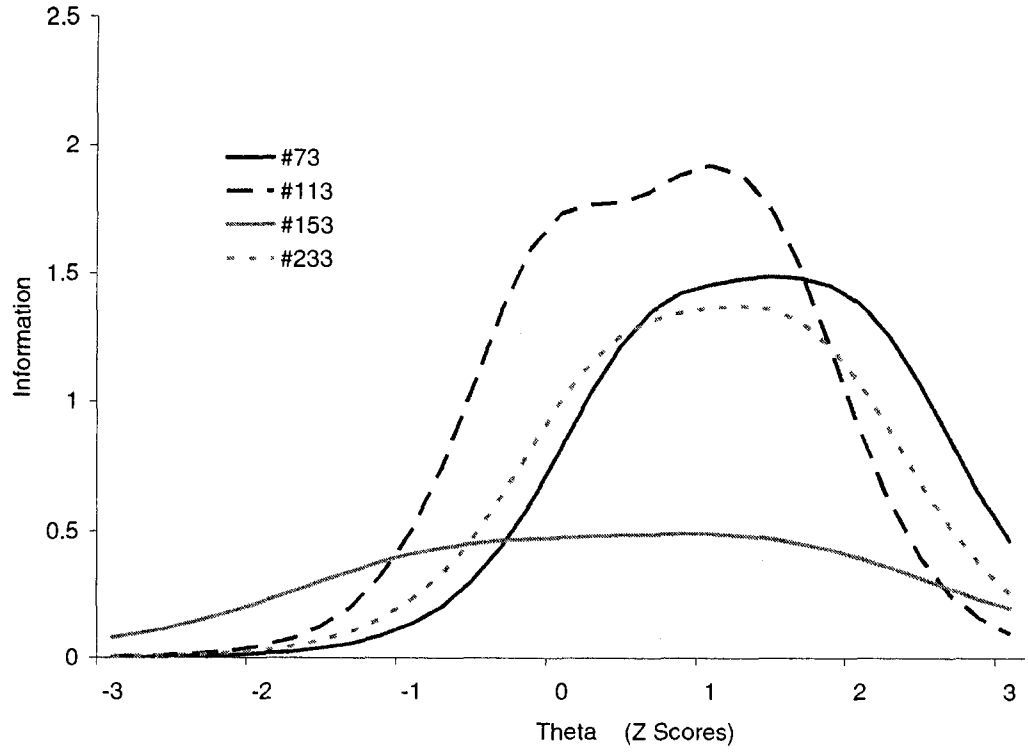


Figure 30. Item information curves for the four items of the ANX-P-R subscale of the PAI based upon responses from a sample of MMT patients (N = 323).

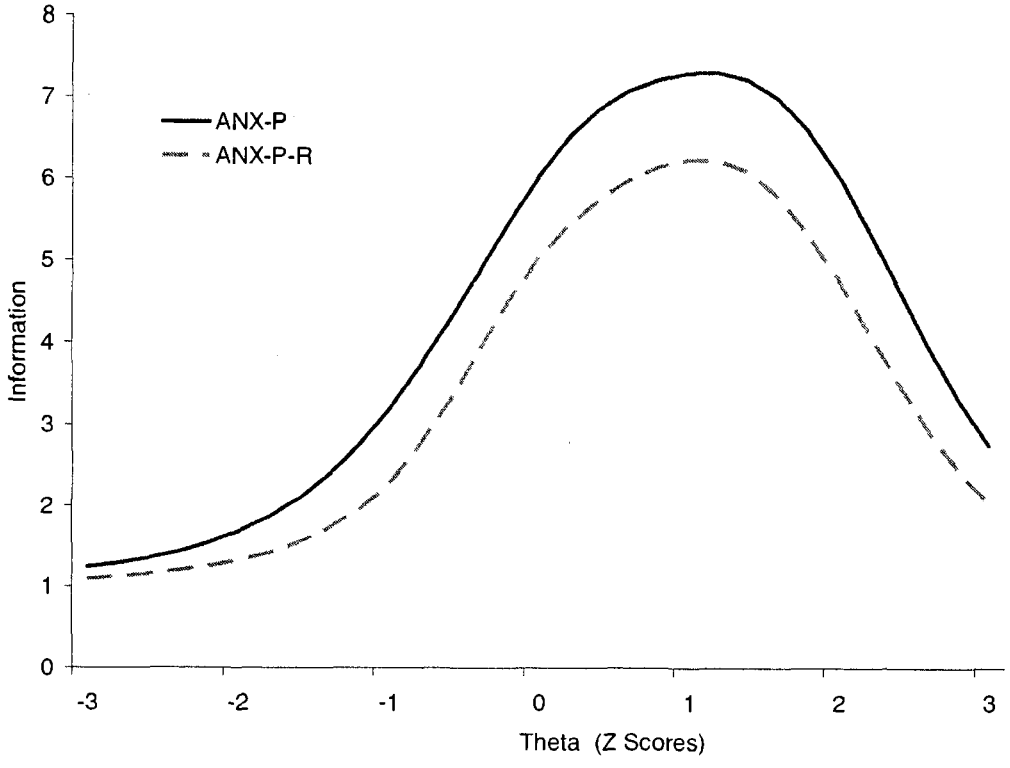
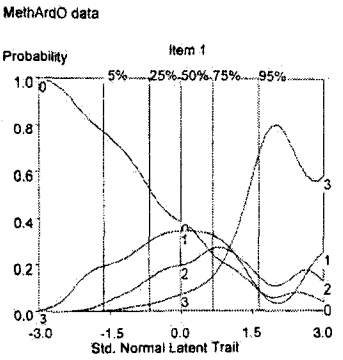
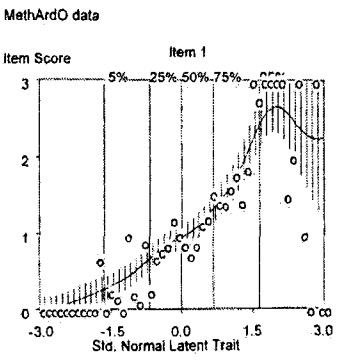


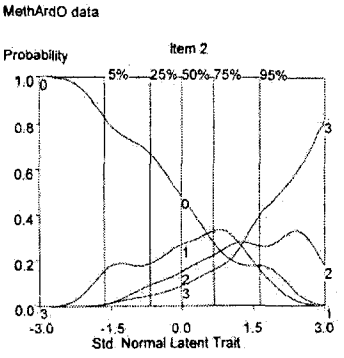
Figure 31. Comparison of the information provided by the eight item ANX-P subscale and the four item ANX-P-R subscale based upon responses from a sample of MMT patients (N = 323).

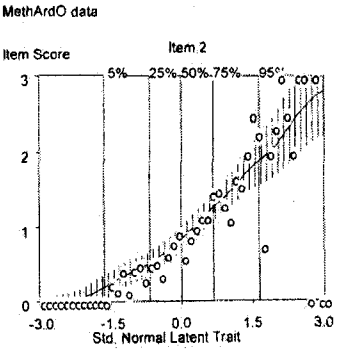
PAI Item #5



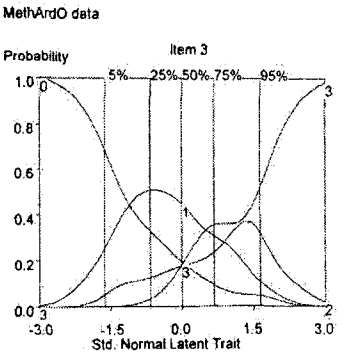


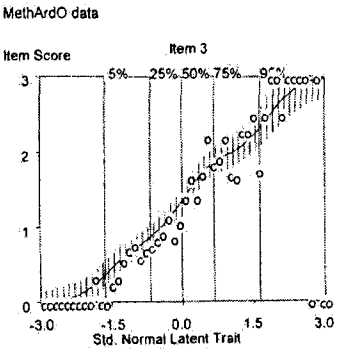
PAI Item #45



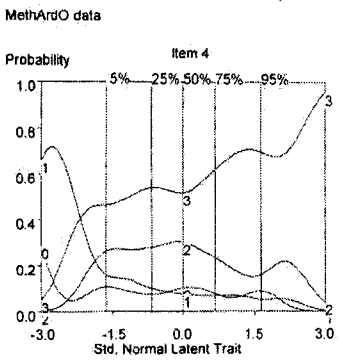


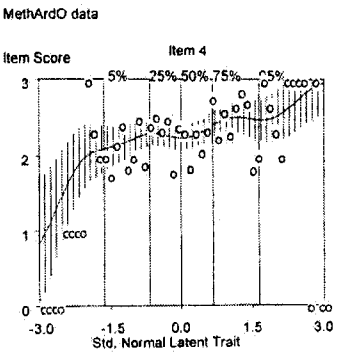
PAI Item #85



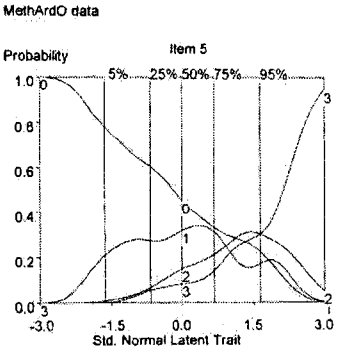


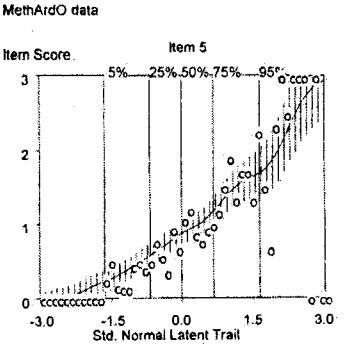
PAI Item #125



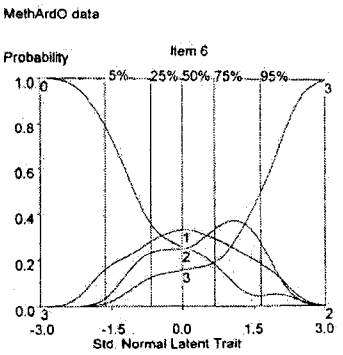


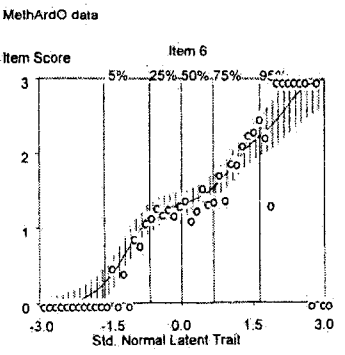
PAI Item #165



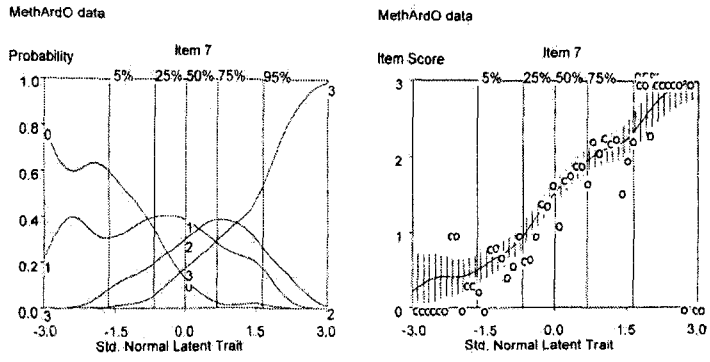


PAI Item #205





PAI Item #245



PAI Item #285

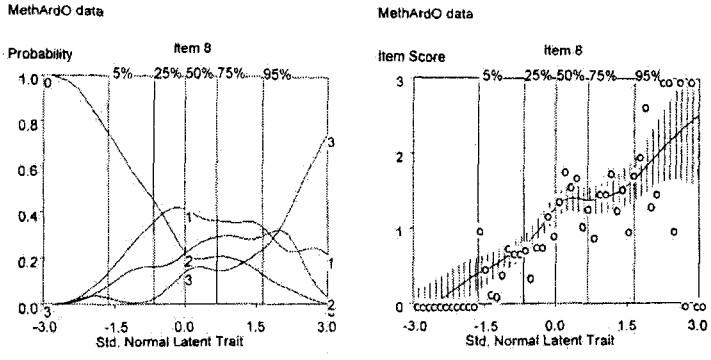


Figure 32. Option and item characteristic curves for the PAI ARD-O subscale employed in a sample of MMT patients ($N = 323$).

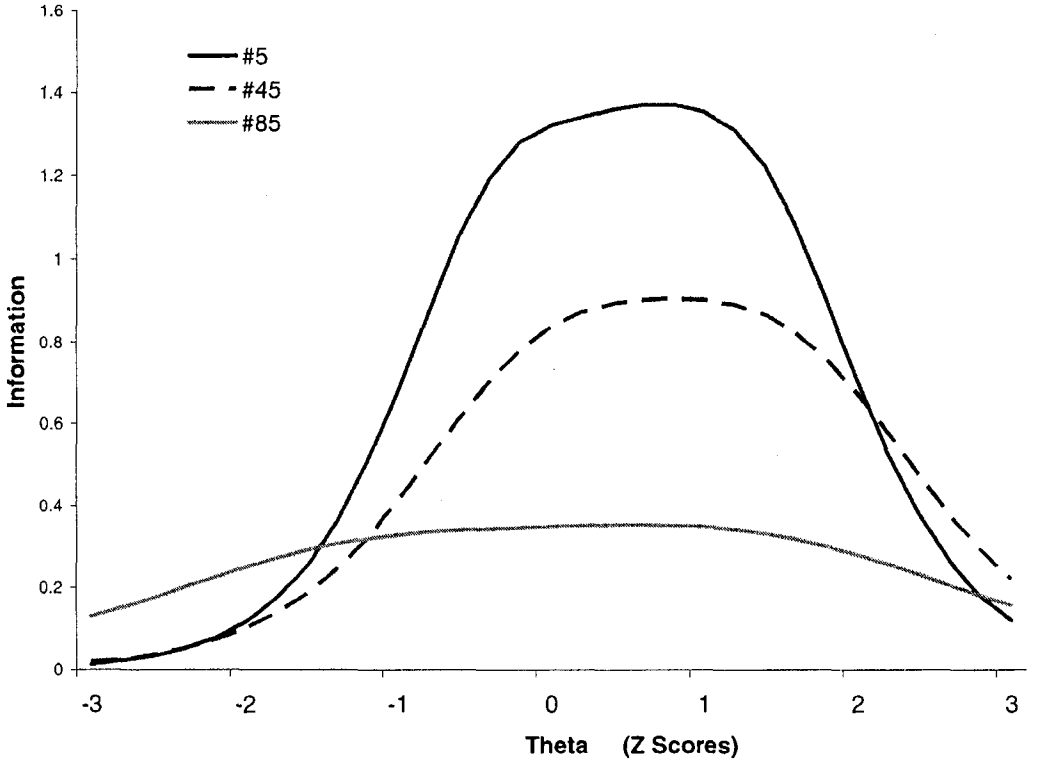
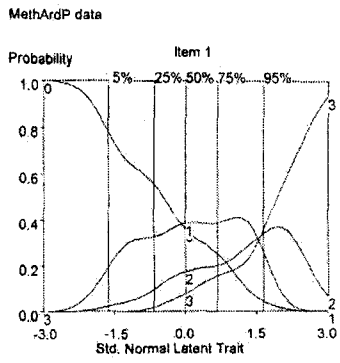
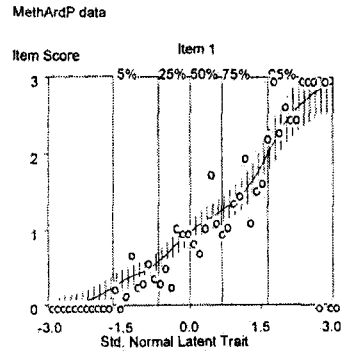


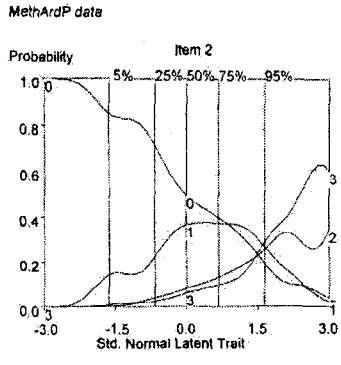
Figure 33. Item information curves for the ARD-O-R subscale in the MMT population (*N* = 323).

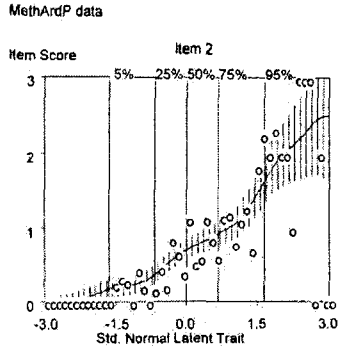
PAI Item #26



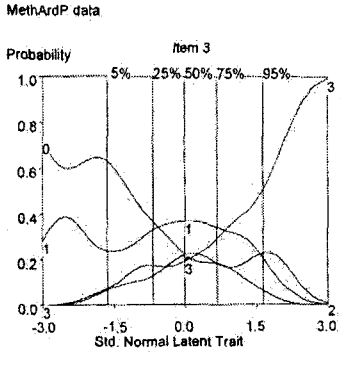


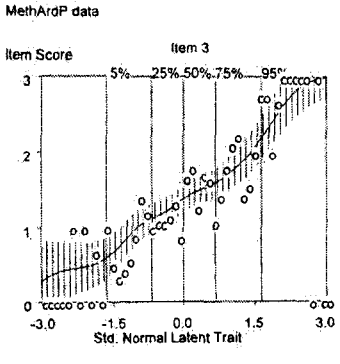
PAI Item #66



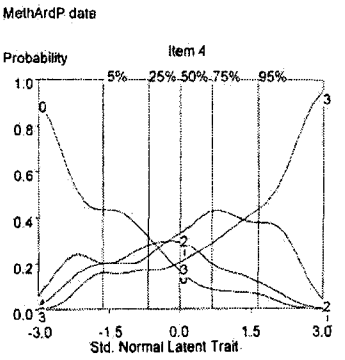


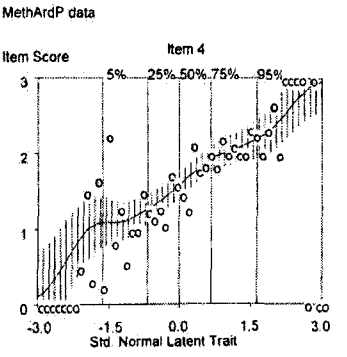
PAI Item #106



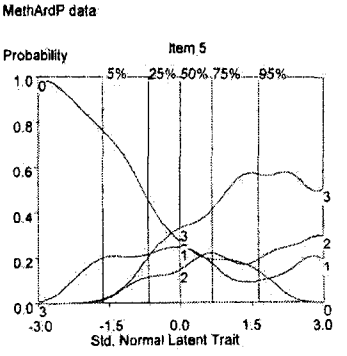


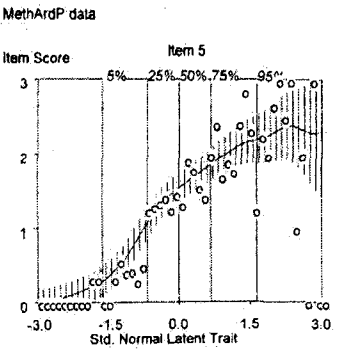
PAI Item #146



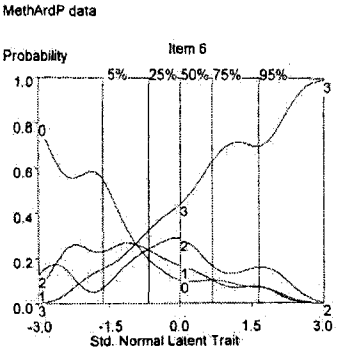


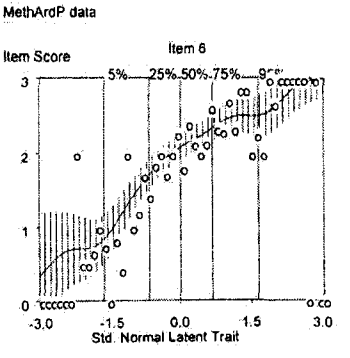
PAI Item #186



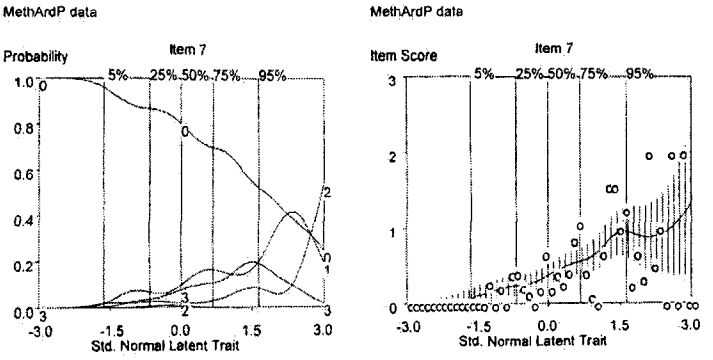


PAI Item #226





PAI Item #266



PAI Item #306

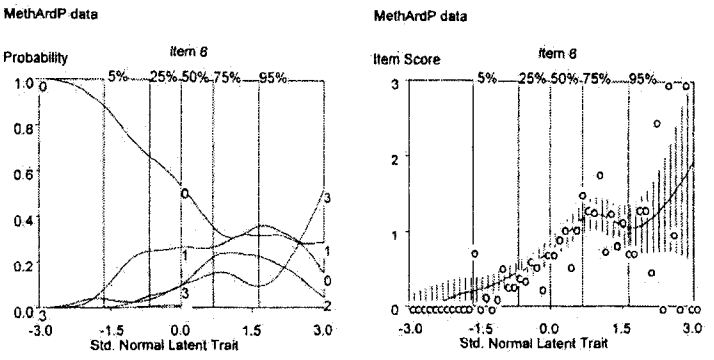


Figure 34. Option and item characteristic curves for the ARD-P subscale of the PAI in a sample of MMT patients ($N = 323$).

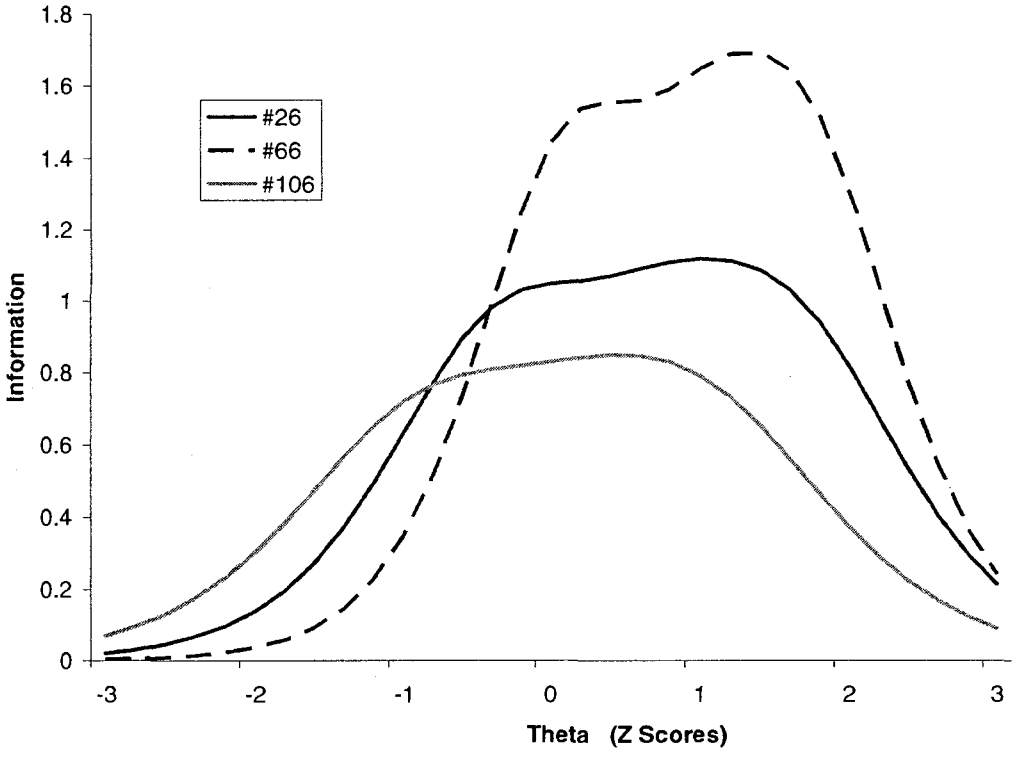


Figure 35. Item information curves for the three items of the ANX-P-R based upon responses to the PAI by a sample of MMT patients ($N = 323$).

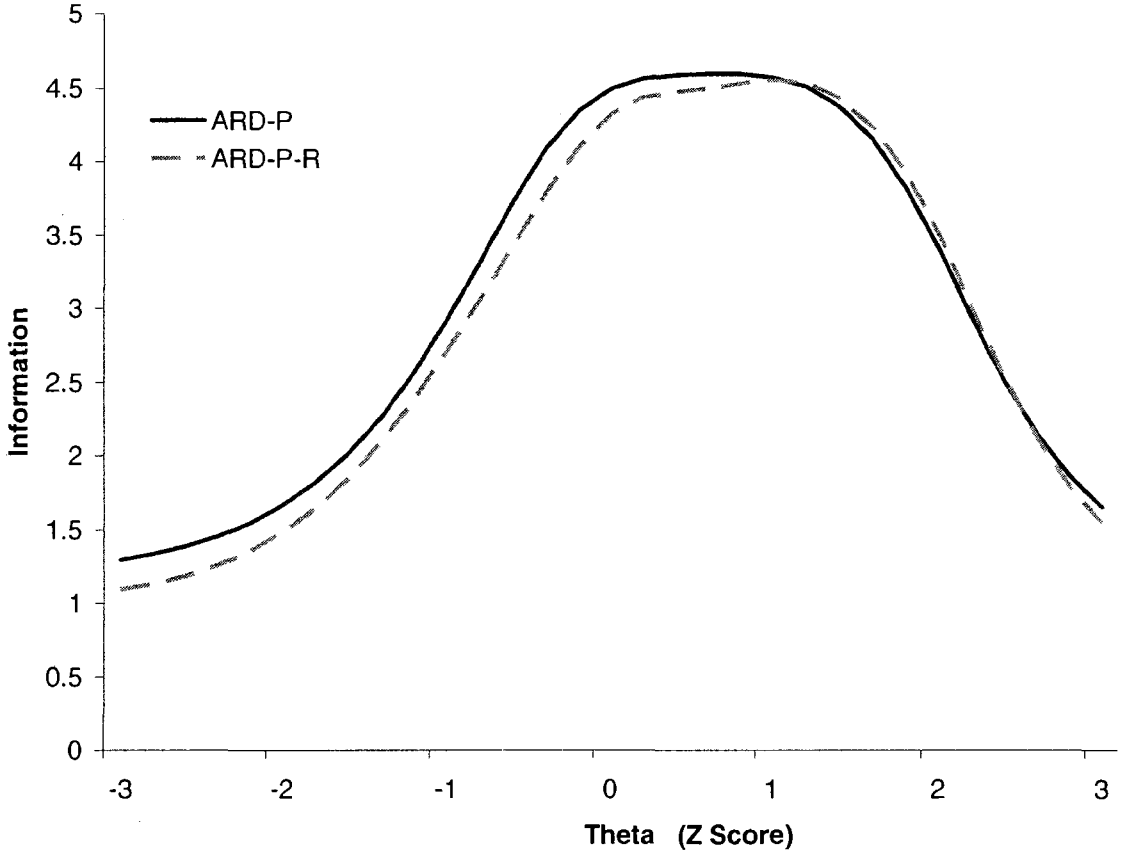
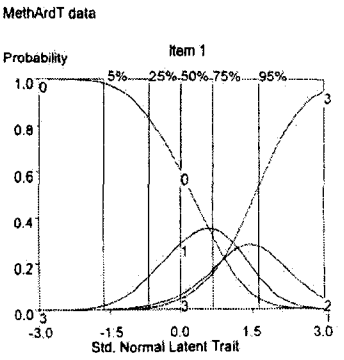
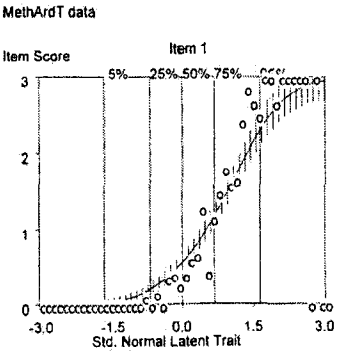


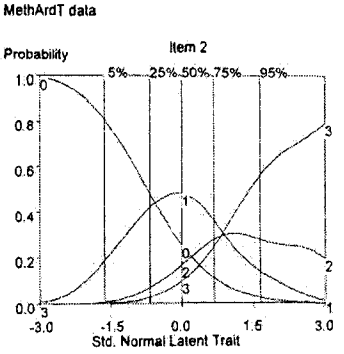
Figure 36. Comparison of total information of the original ARD-P (8 items) and the ARD-P-S subscale (3 items) in a sample of MMT patients ($N = 323$).

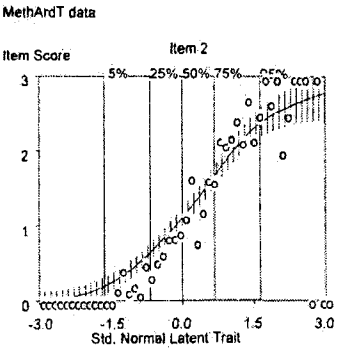
PAI Item #34



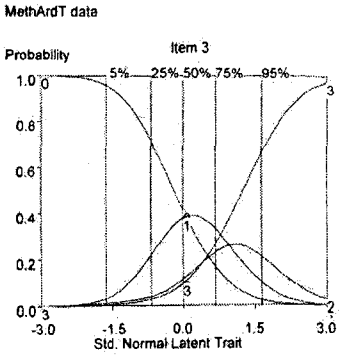


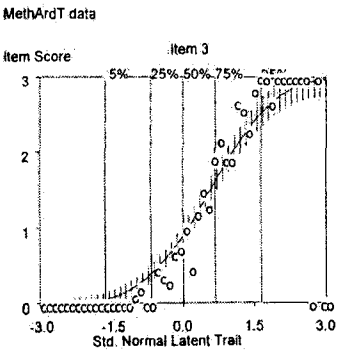
PAI Item #74



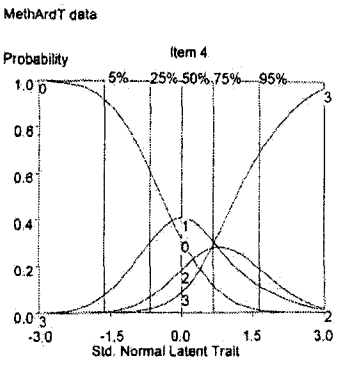


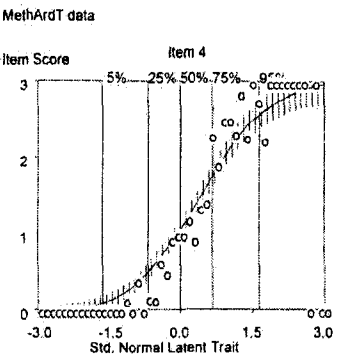
PAI Item #114



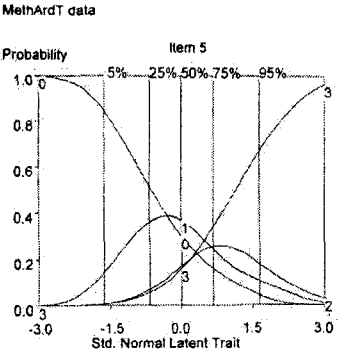


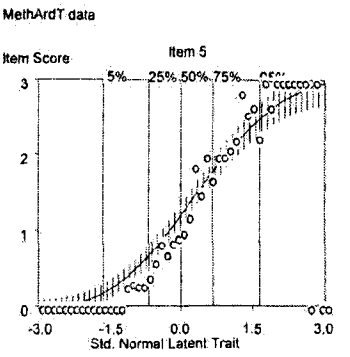
PAI Item #154



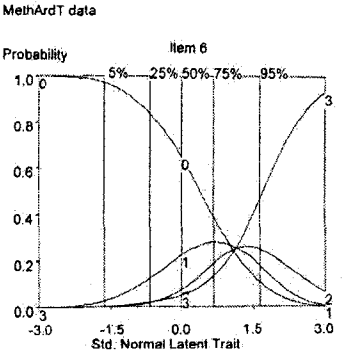


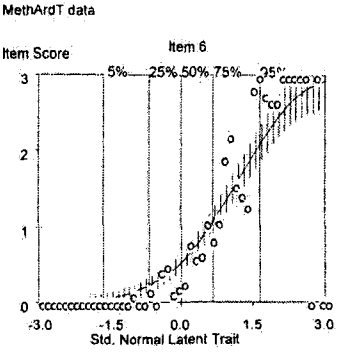
PAI Item #194



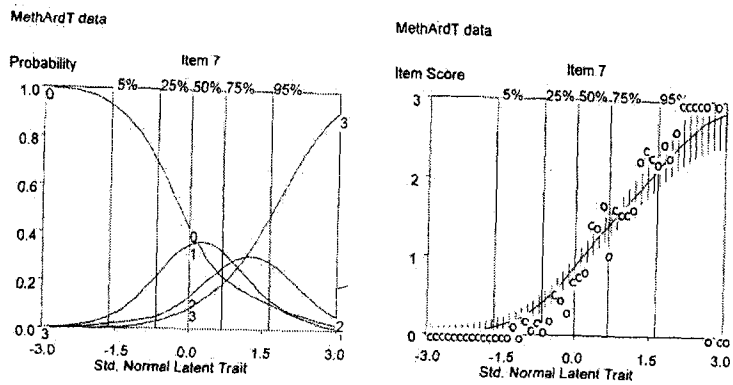


PAI Item #234





PAI Item #274



PAI Item #314

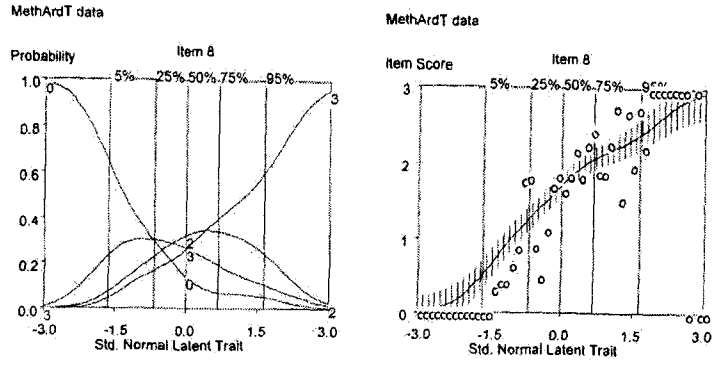


Figure 37. Option and item information curves for the ARD-T subscale in a sample of MMT patients ($N = 323$).

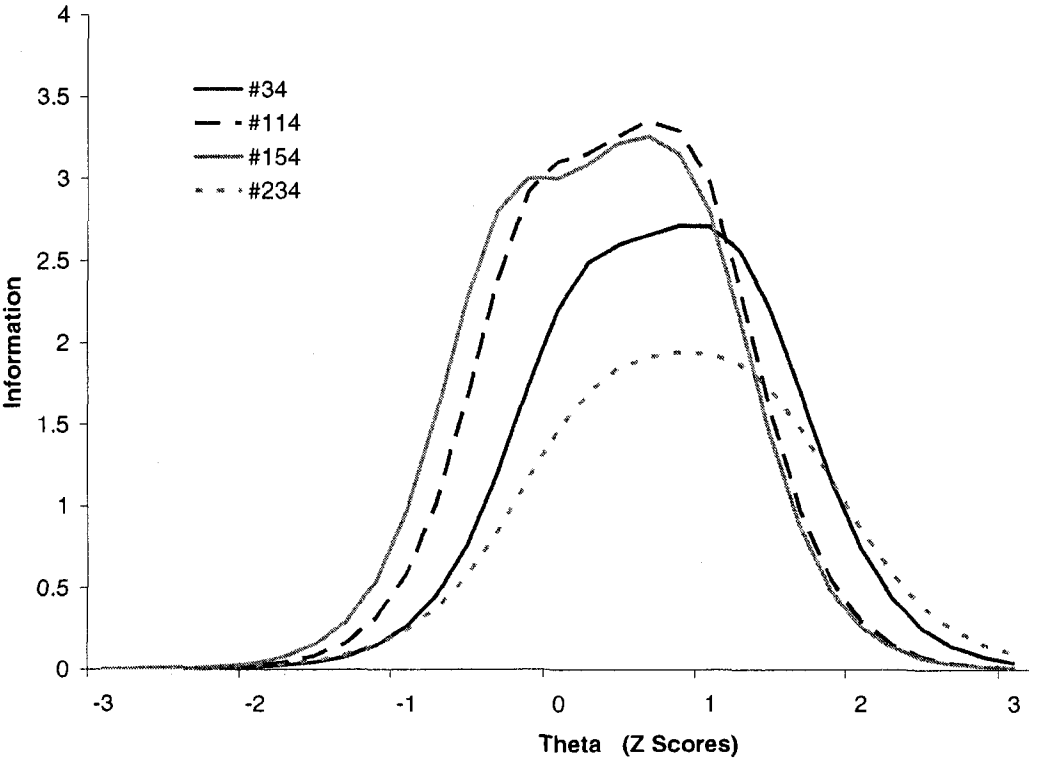


Figure 38. Item information curves for the four items of the ARD-T-R subscale based upon responses by a sample of MMT patients ($N = 323$).

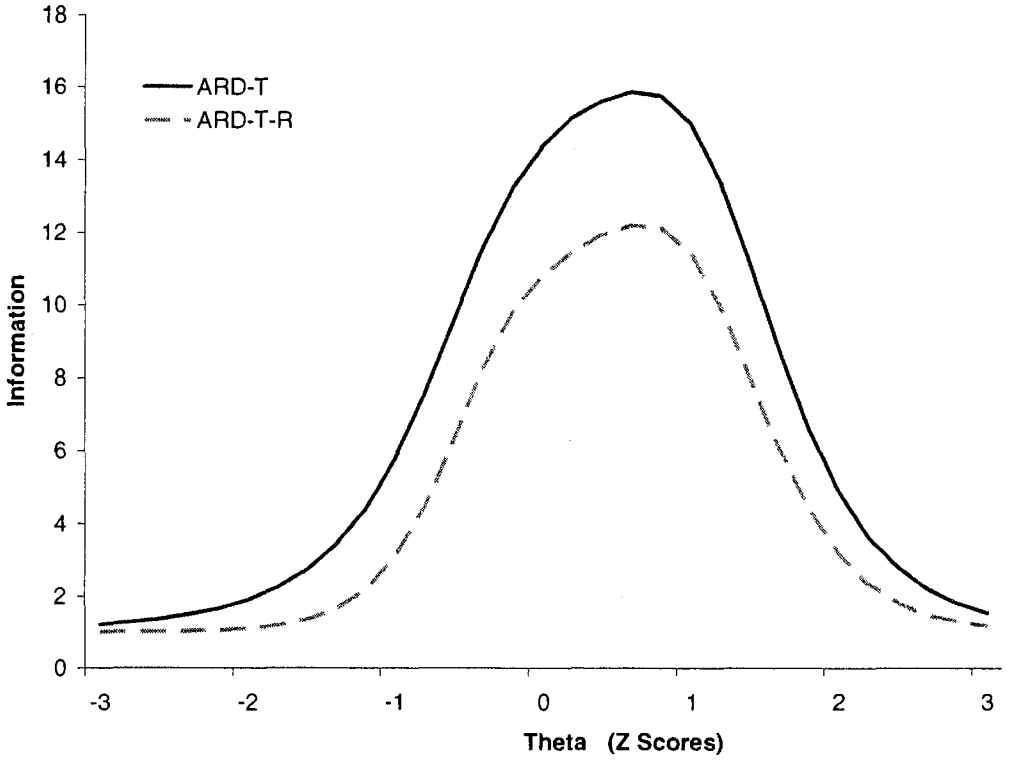
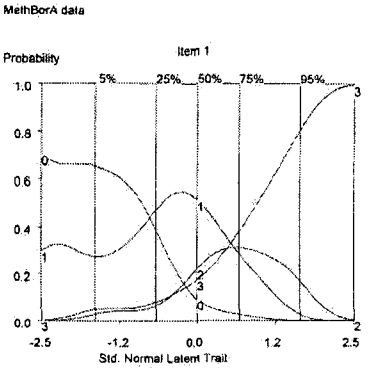
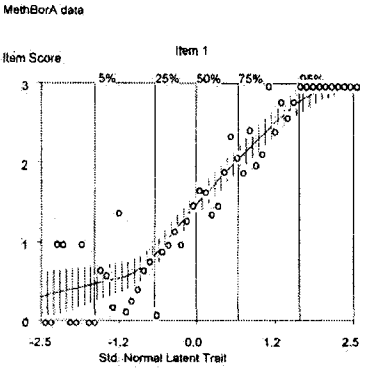


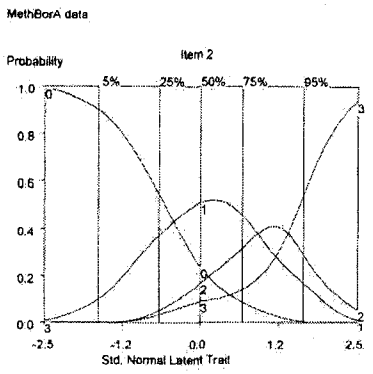
Figure 39. Total subscale information curves comparing the original ARD-T to the ARD-T-R based upon responses from a sample of MMT patients ($N = 323$).

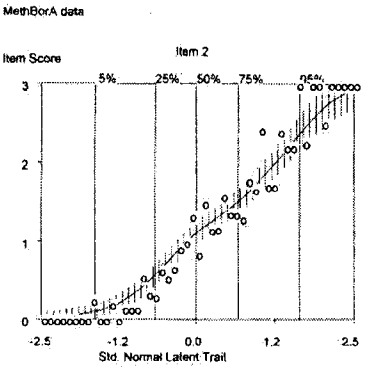
PAI Item #14



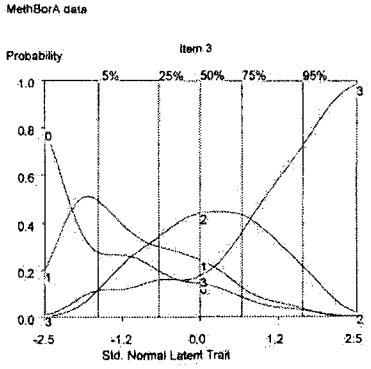


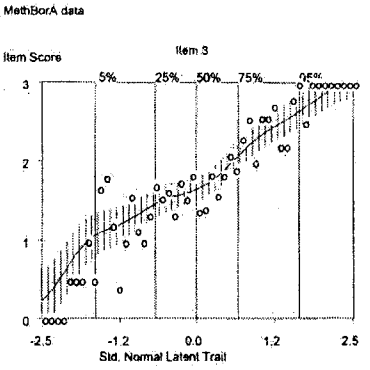
PAI Item #54



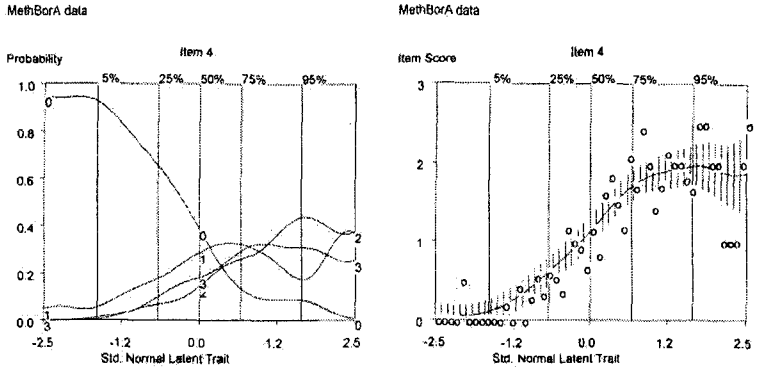


PAI Item #94

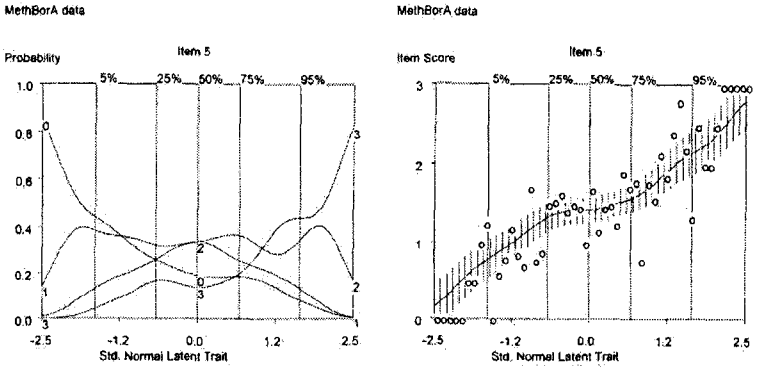




PAI Item #134



PAI Item #174



PAI Item #214

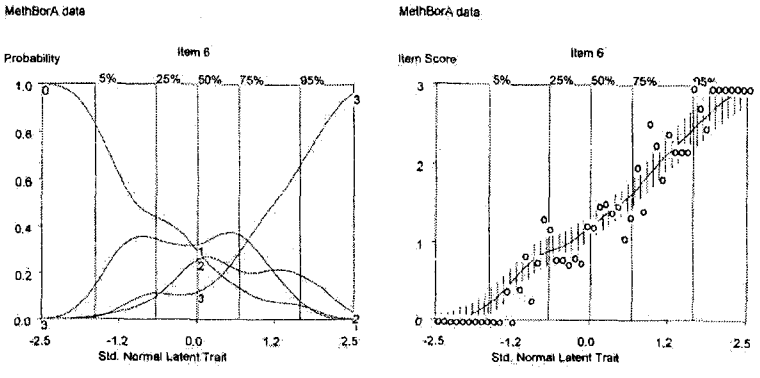


Figure 40. Option and item characteristic curves for the six items of the PAI subscale BOR-A when administered to a sample of MMT patients ($N = 323$).

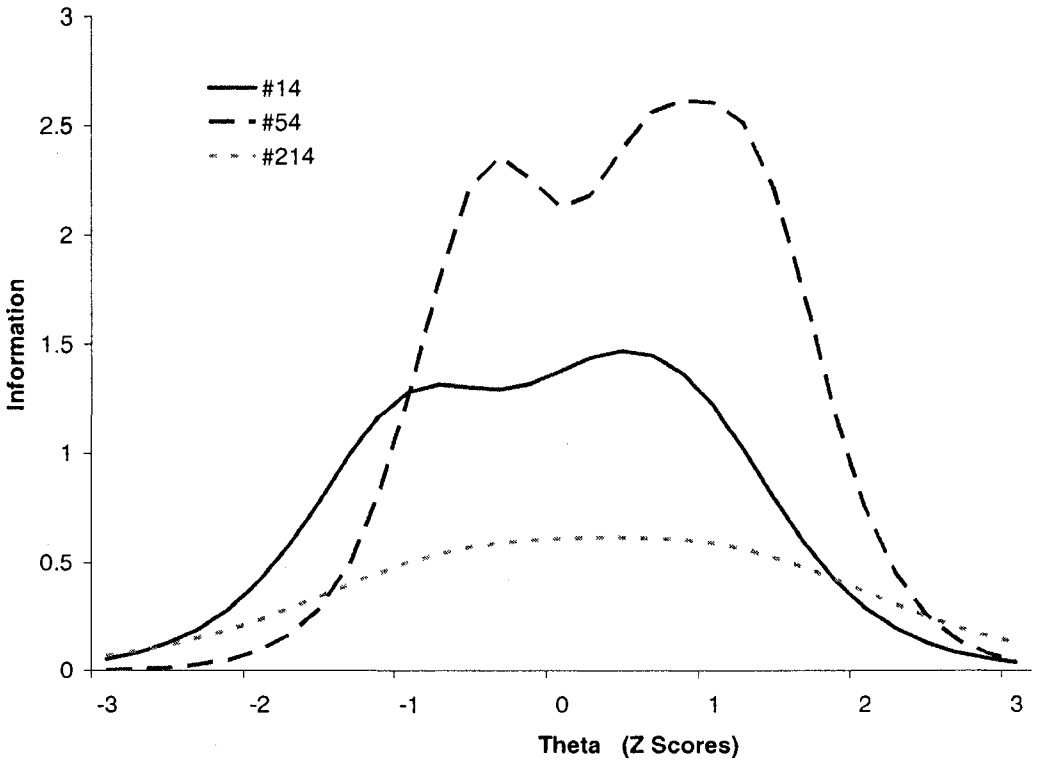


Figure 41. Item information curves for the three items of the BOR-A-R subscale of the PAI based upon responses to PAI by a sample of MMT patients ($N = 323$).

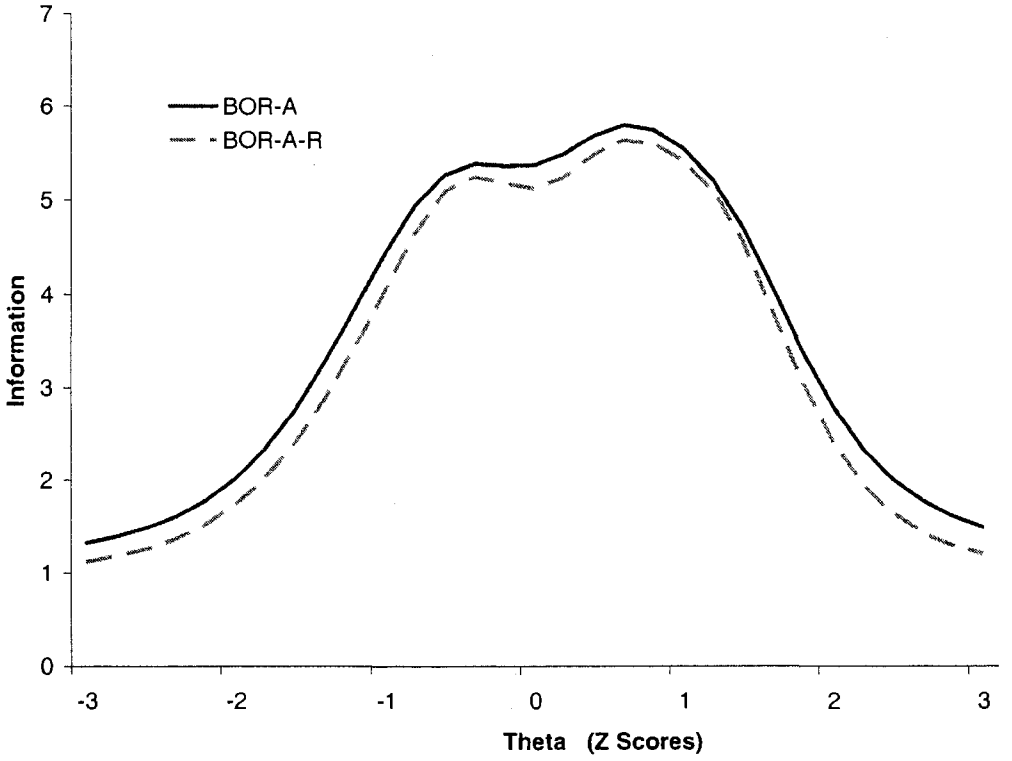
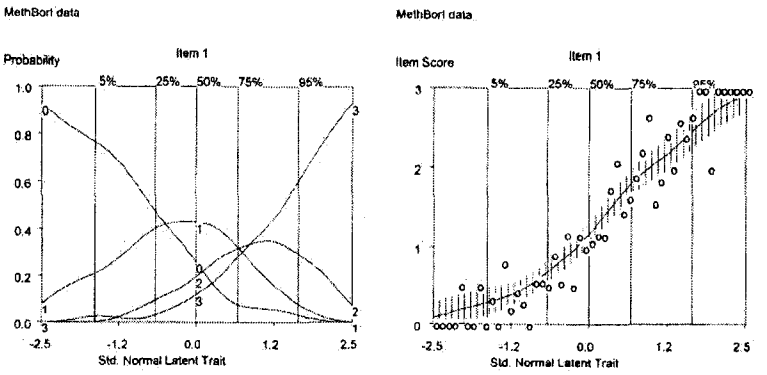
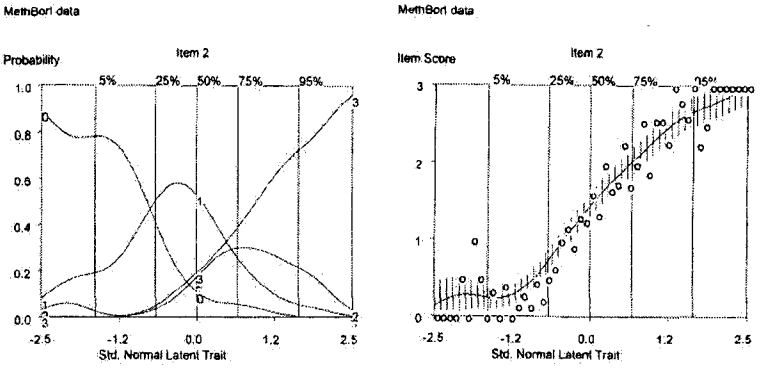


Figure 42. Total scale information curves comparing the original 8 item BOR-A subscale and the 4 item revised version ($N = 323$).

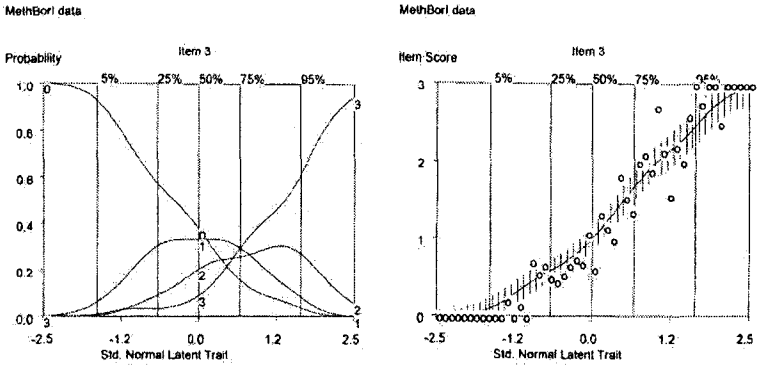
PAI Item #17



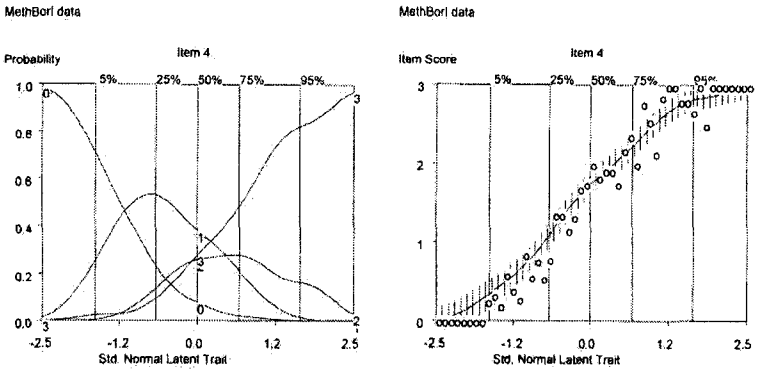
PAI Item #57



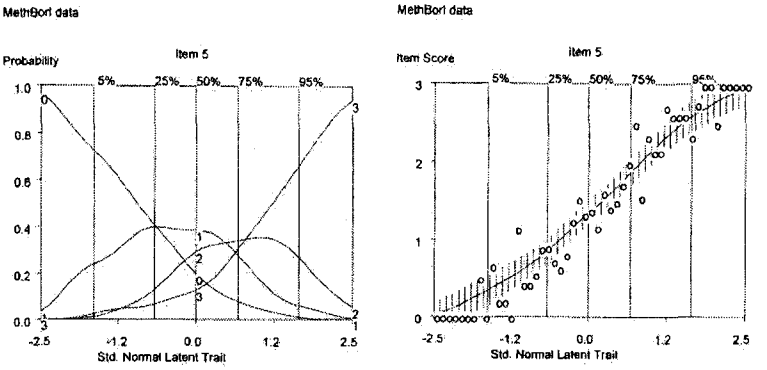
PAI Item #97



PAI Item #137



PAI Item #177



PAI Item #214

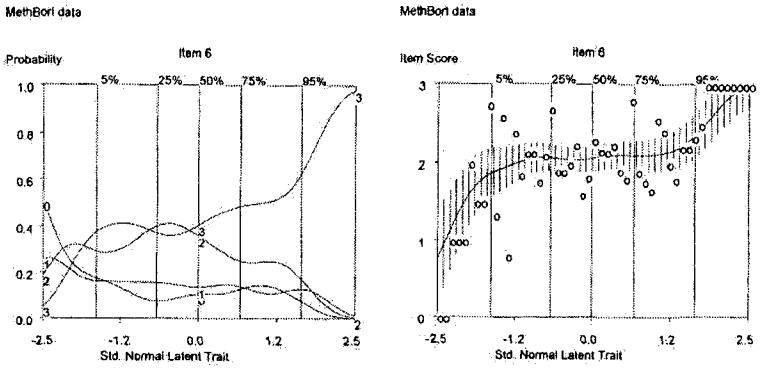


Figure 43. Option and item characteristic curves for the BOR-I subscale of the PAI based upon responses to the PAI by a sample of MMT patients ($N = 323$).

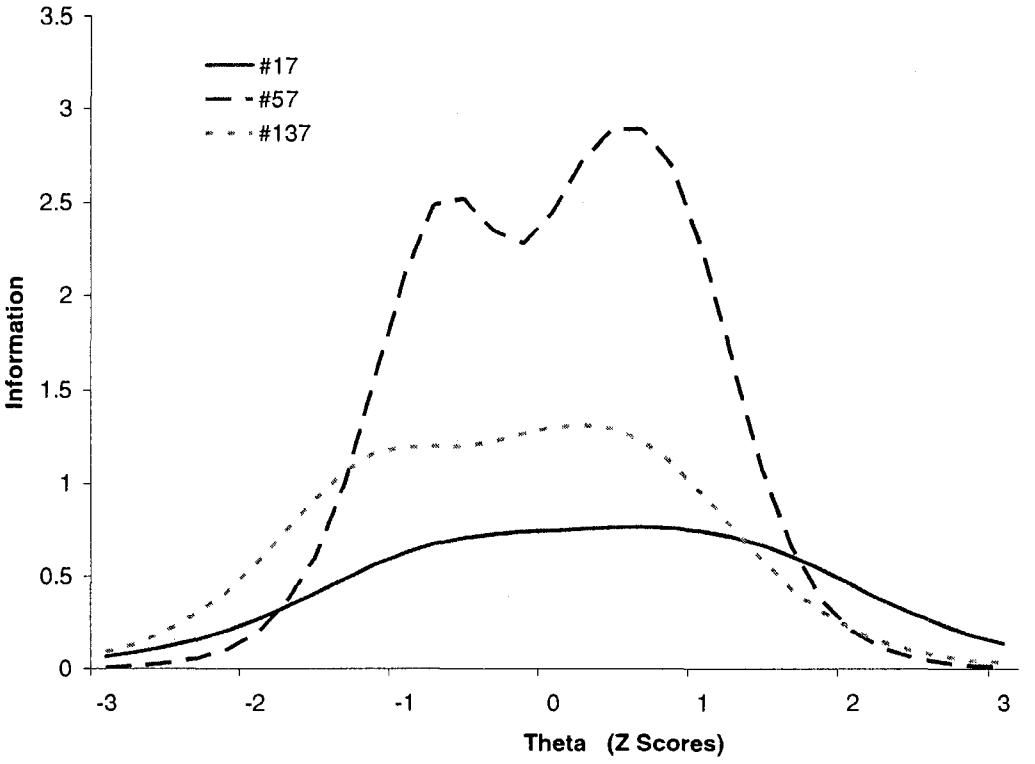
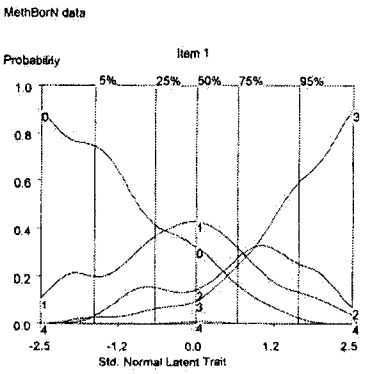
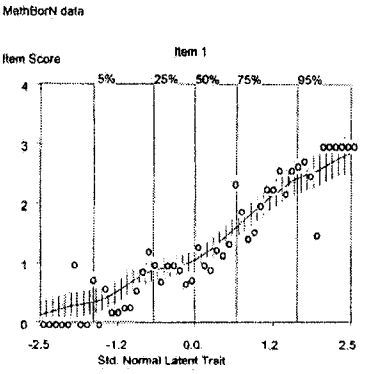


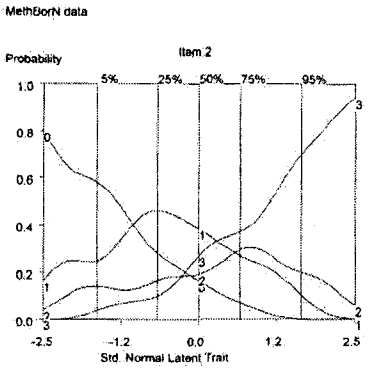
Figure 44. Item information curves for the three items of the BOR-I-R as administered to a sample of MMT patients ($N = 323$).

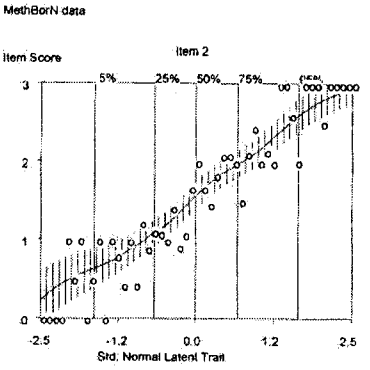
PAI Item #19



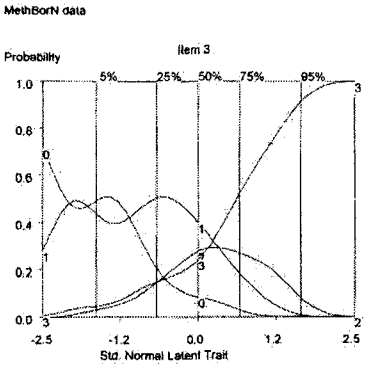


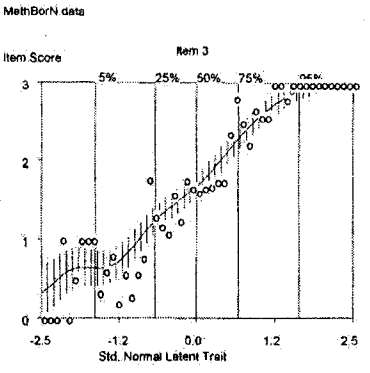
PAI Item #59



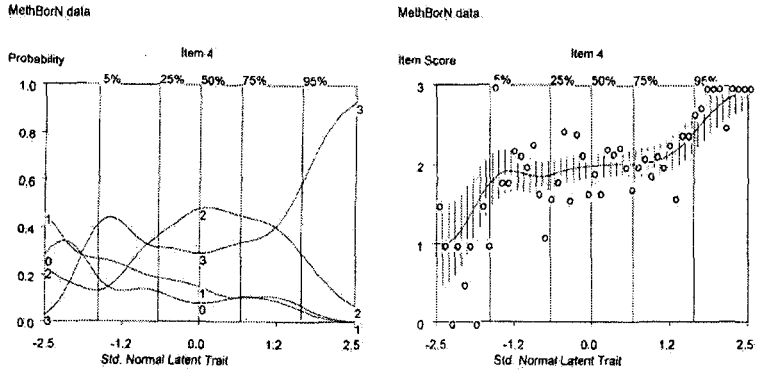


PAI Item #99

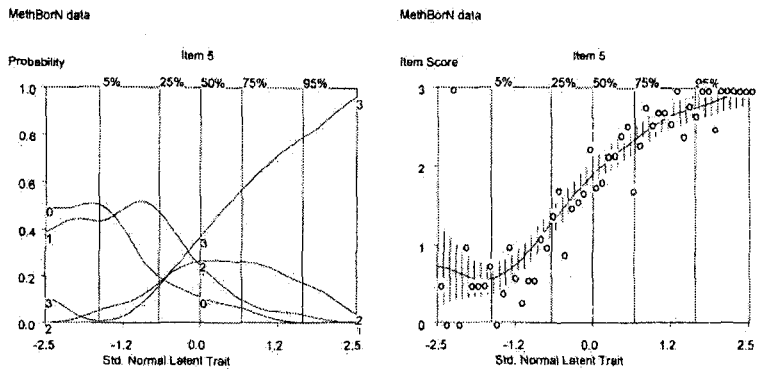




PAI Item #139



PAI Item #179



PAI Item #219

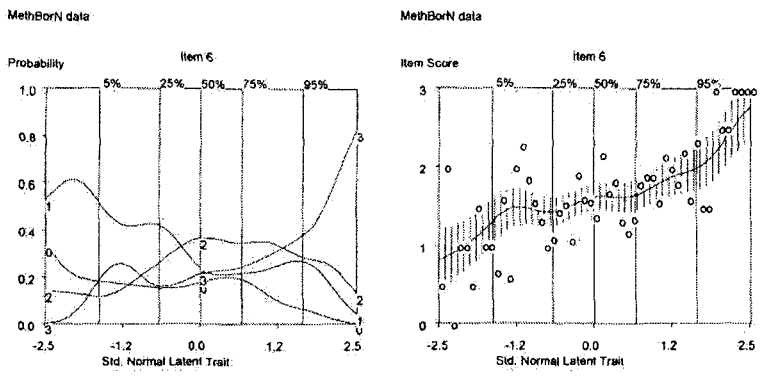


Figure 45. Option and item characteristic curves for the BOR-N subscale based upon responses to the PAI by a sample of MMT patients (N = 323).

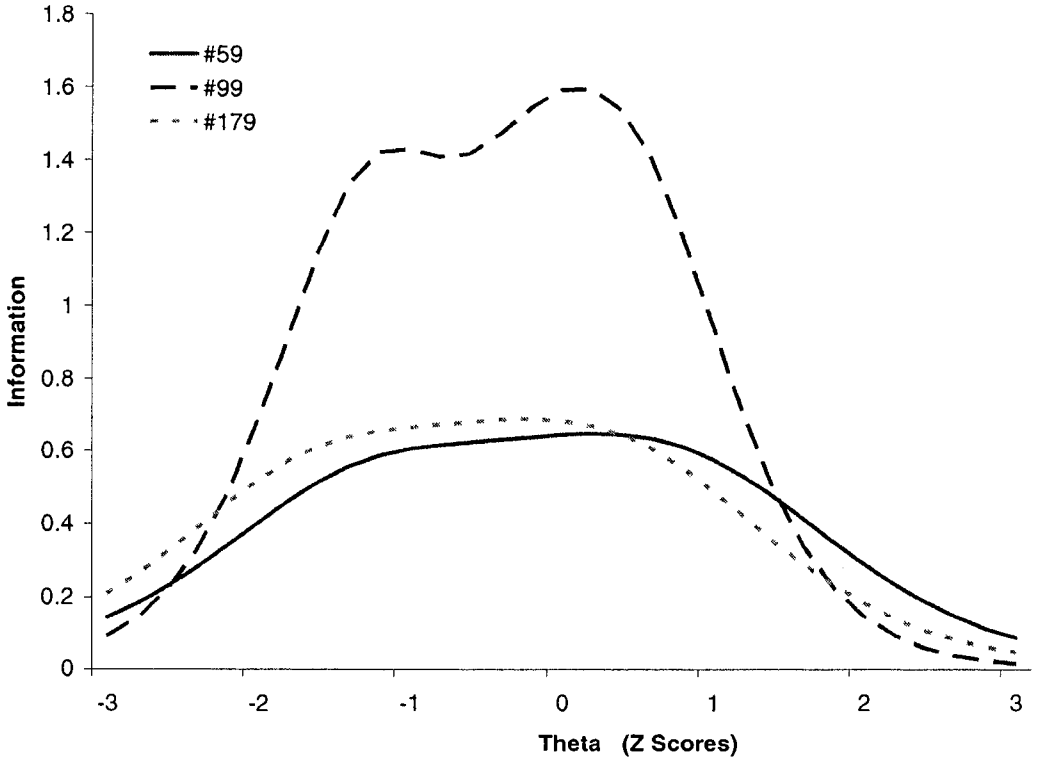
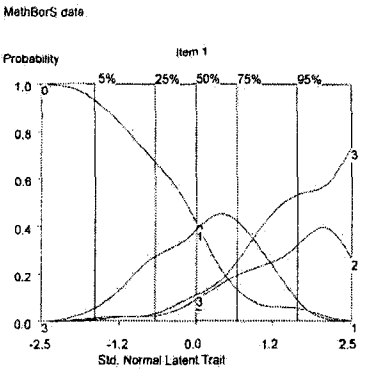
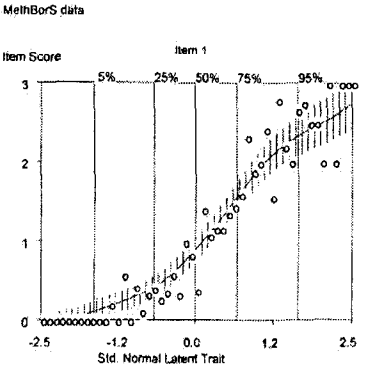


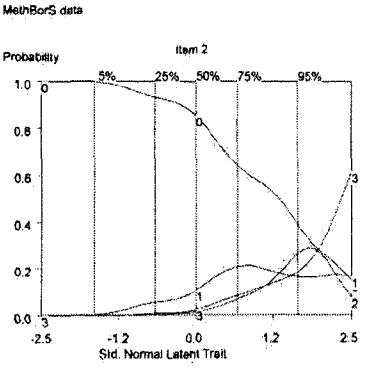
Figure 46. Item information curves for the three items of the BOR-N-R subscale based upon responses from a sample of MMT patients ($N = 323$).

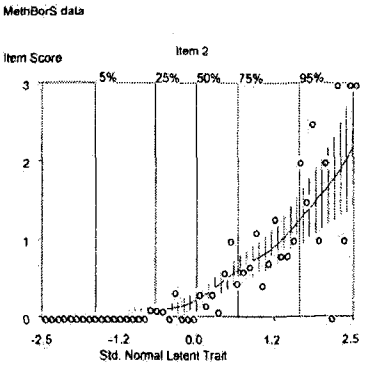
PAI Item# 143



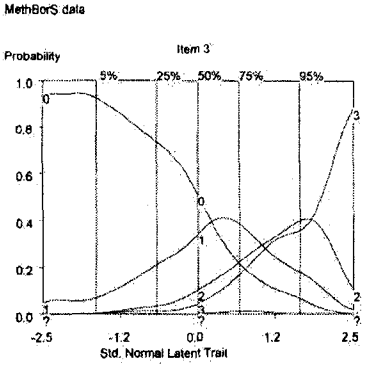


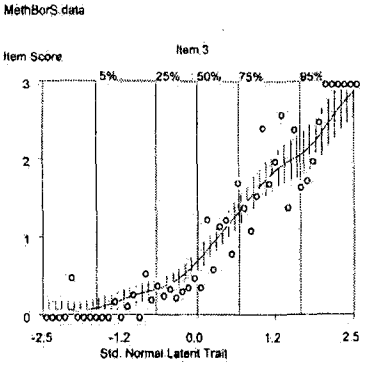
PAI Item#183



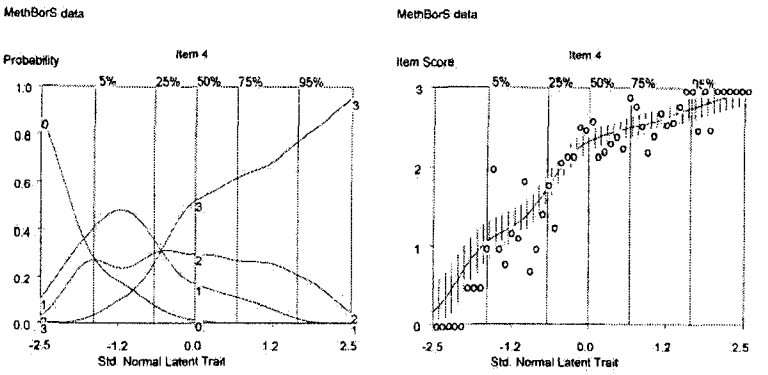


PAI Item#223

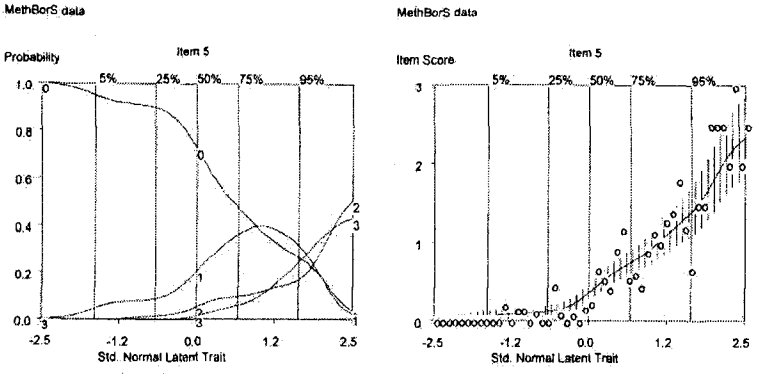




PAI Item#263



PAI Item#303



PAI Item#343

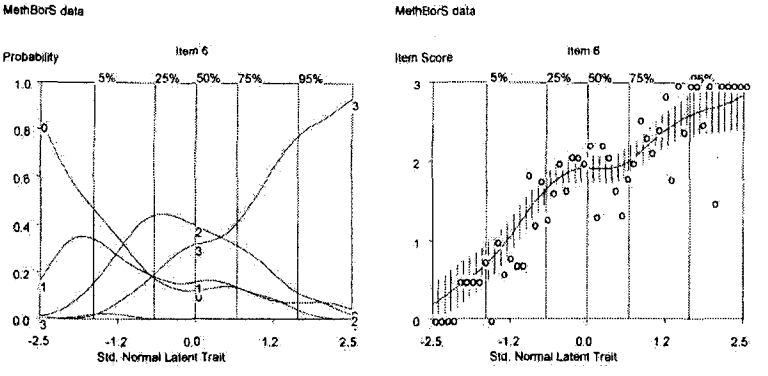


Figure 47. Option and item characteristic curves for the BOR-S subscale based upon responses from a sample of MMT patients (N = 323).

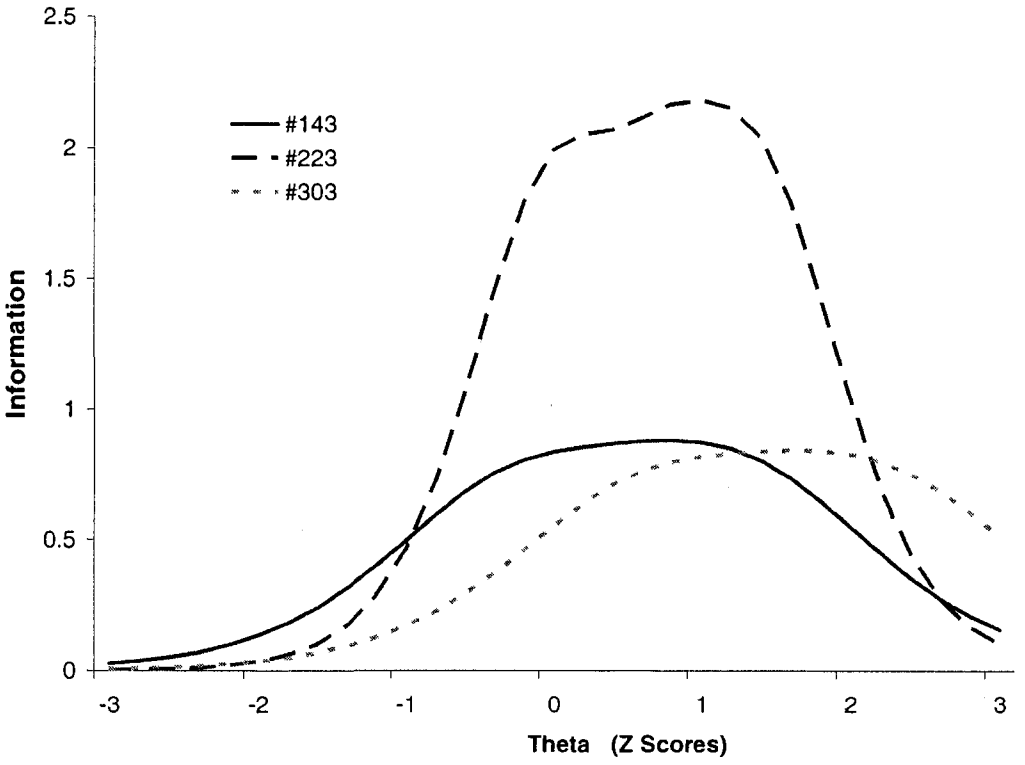
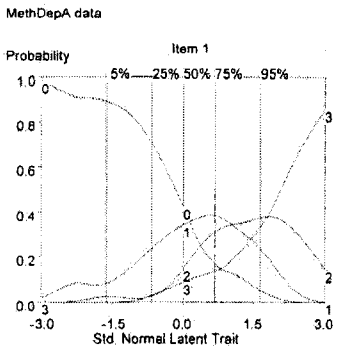
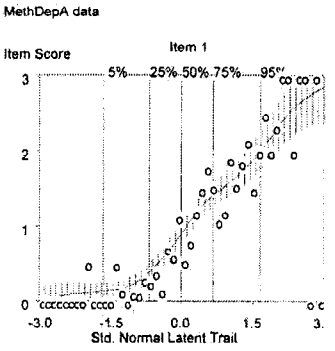


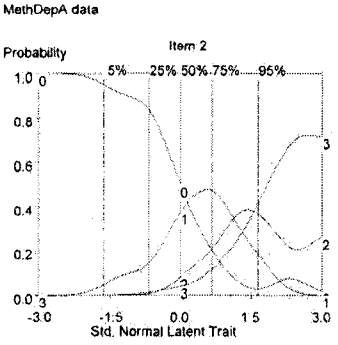
Figure 48. Information curves for the three items of the BOR-S-R subscale based upon responses from a sample of MMT (N = 323).

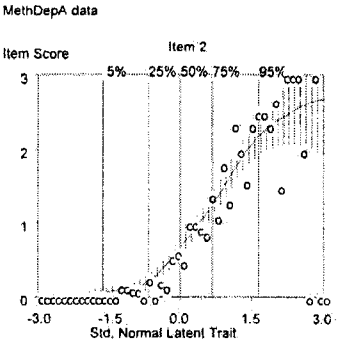
PAI Item #6



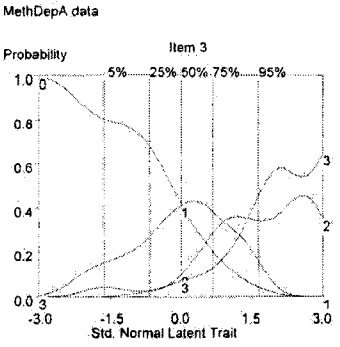


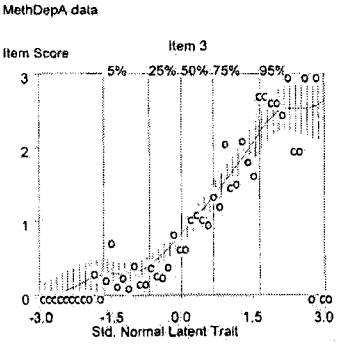
PAI Item #46



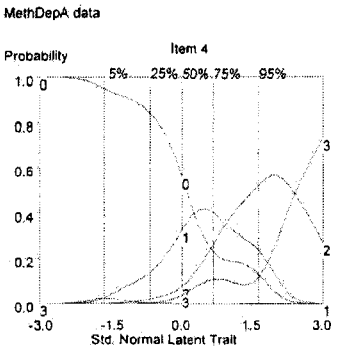


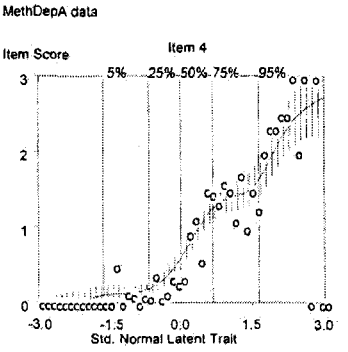
PAI Item #86



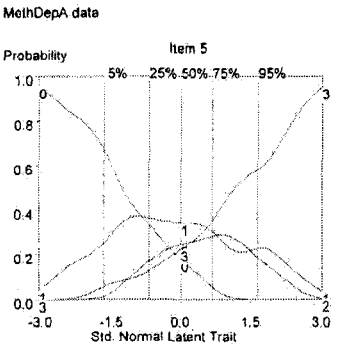


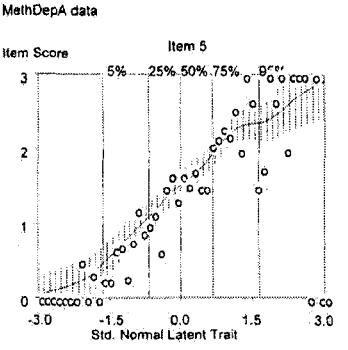
PAI Item #126



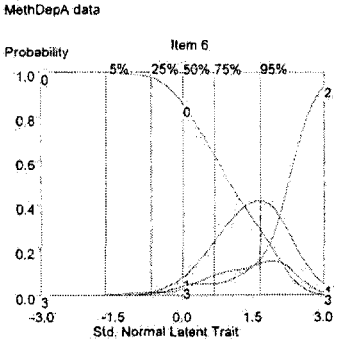


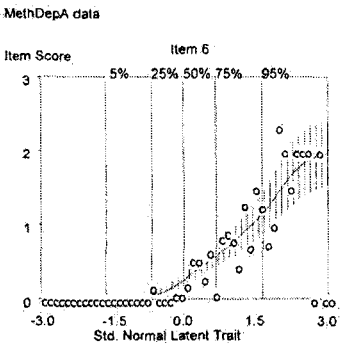
PAI Item #166



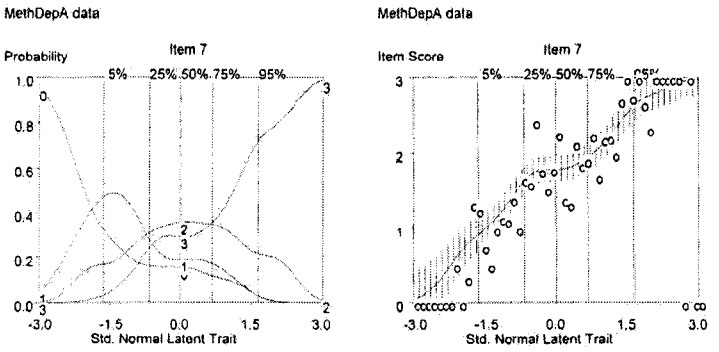


PAI Item #206





PAI Item #246



PAI Item #286

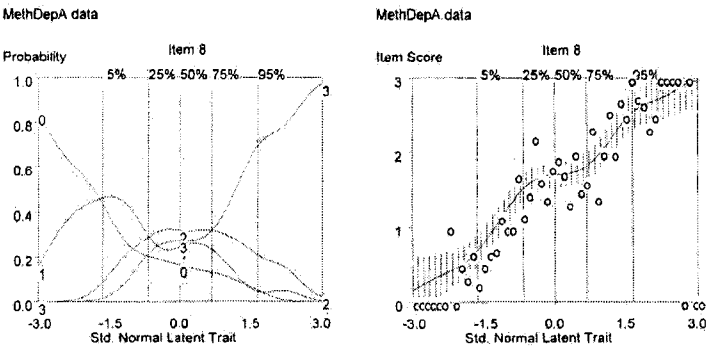


Figure 49. Option and item characteristic curves for the DEP-A subscale based upon responses from a sample of MMT patients ($N = 323$).

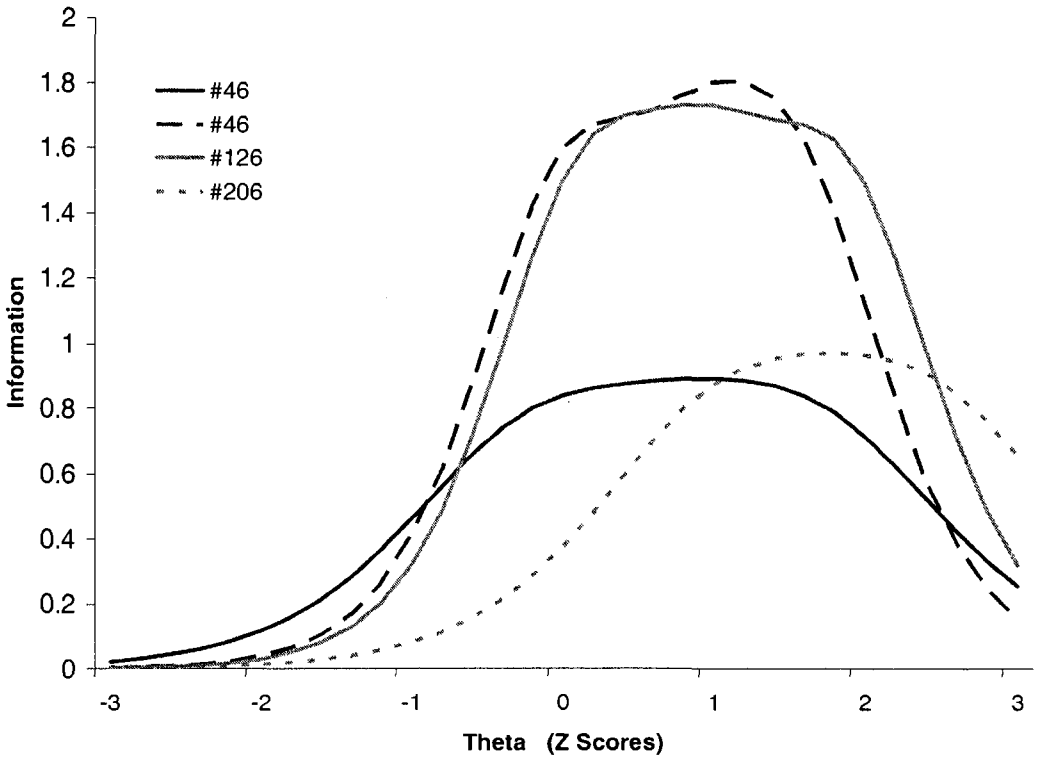


Figure 50. Item information curves for the four items of the DEP-A-R subscale based upon the responses of a sample of MMT patients (N = 323).

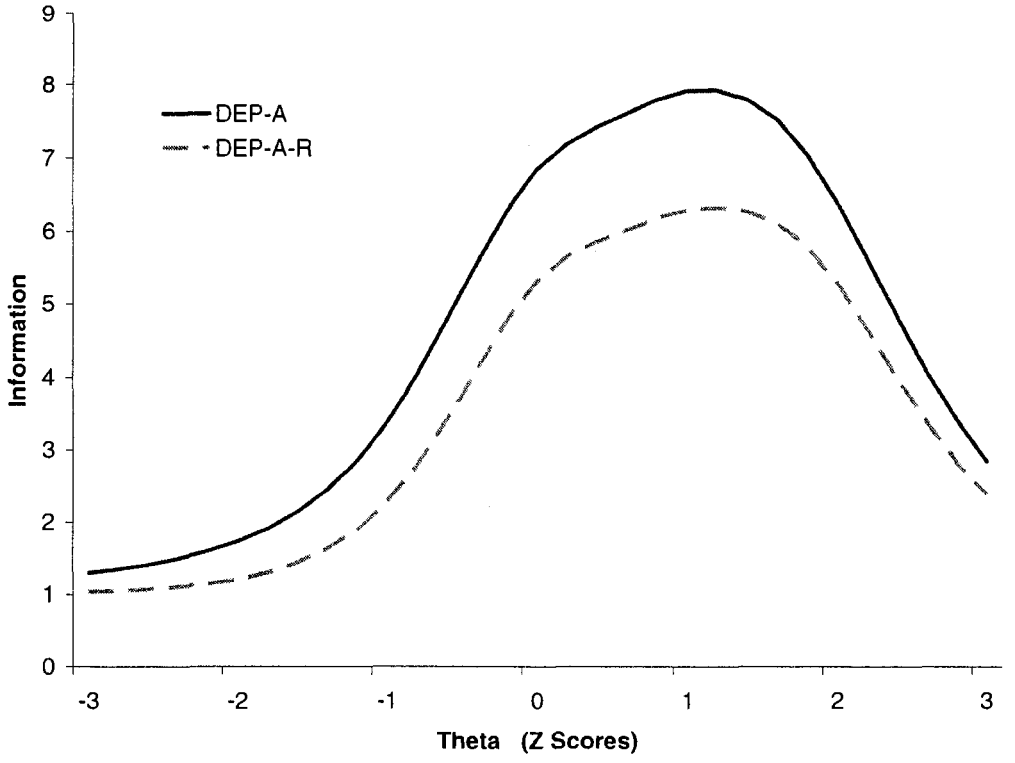
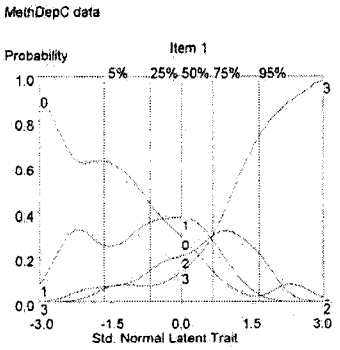
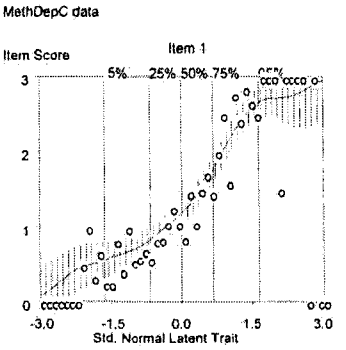


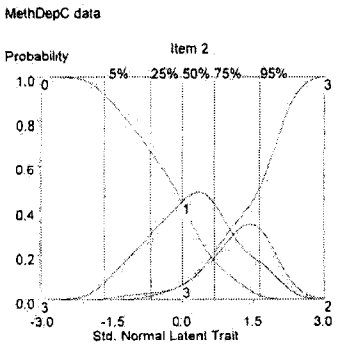
Figure 51. Comparison of total information of the 8 item DEP-A subscale and the four item DEP-A-R subscale based upon responses of a sample of MMT patients (N = 323).

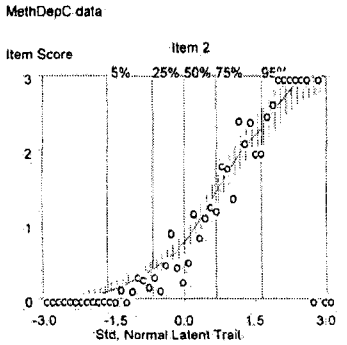
PAI Item #27



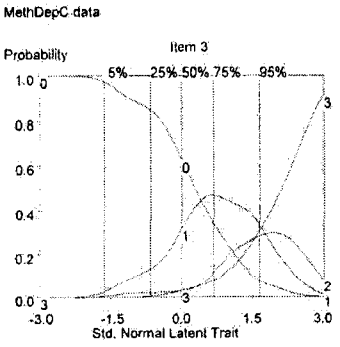


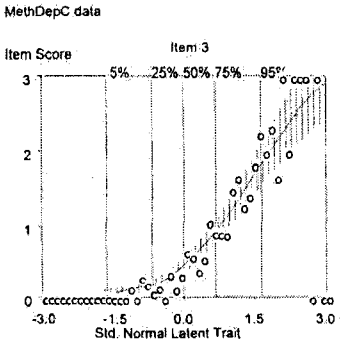
PAI Item #67



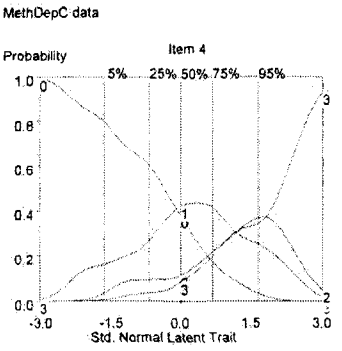


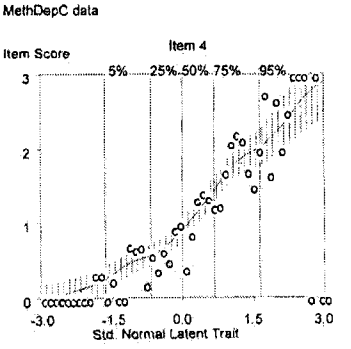
PAI Item #107



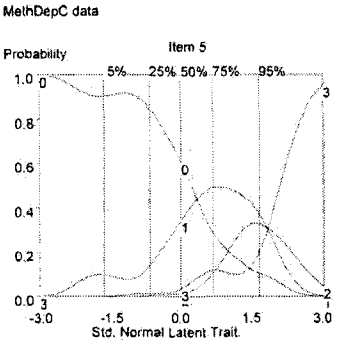


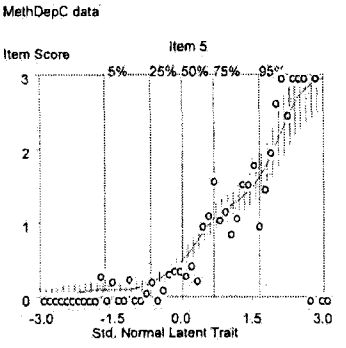
PAI Item #147



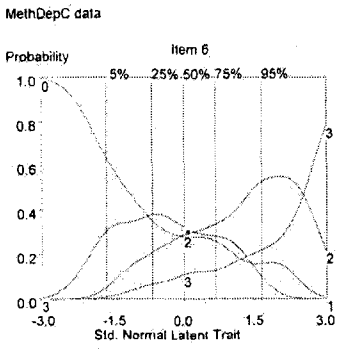


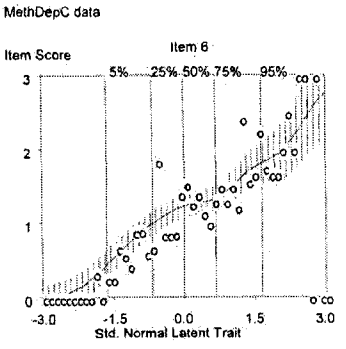
PAI Item #187



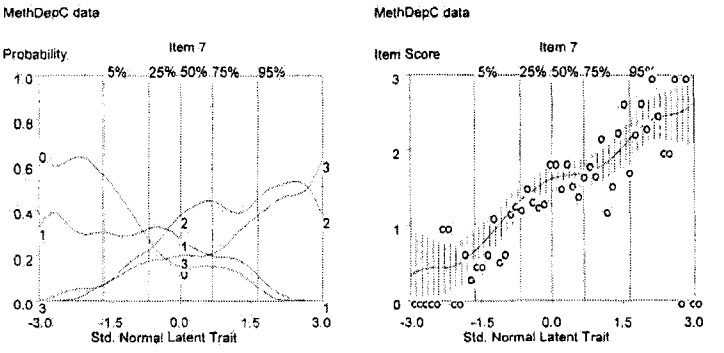


PAI Item #227





PAI Item #267



PAI Item #307

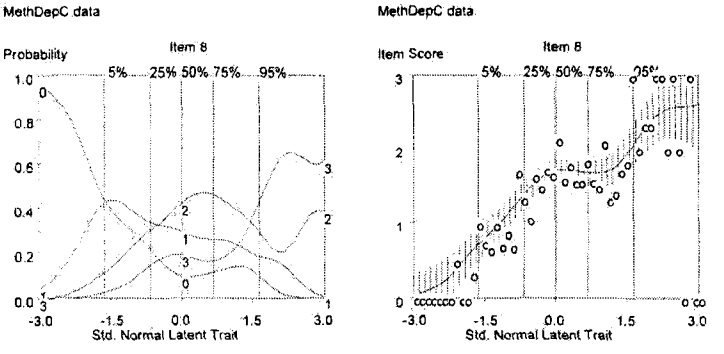


Figure 52. Option and item characteristic curves of the PAI subscale DEP-C as plotted by Testgraf based upon a sample of MMT patients ($N = 323$).

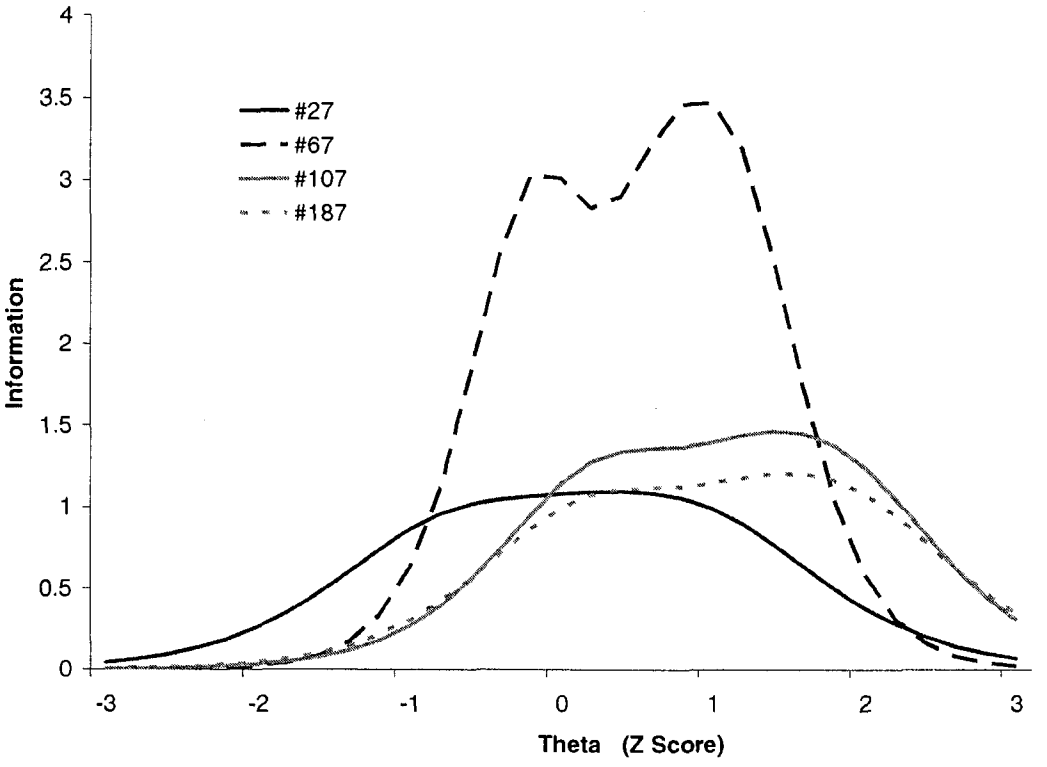


Figure 53. Item information curves for the DEP-C-R subscale based upon response from a sample of MMT patients (N = 323).

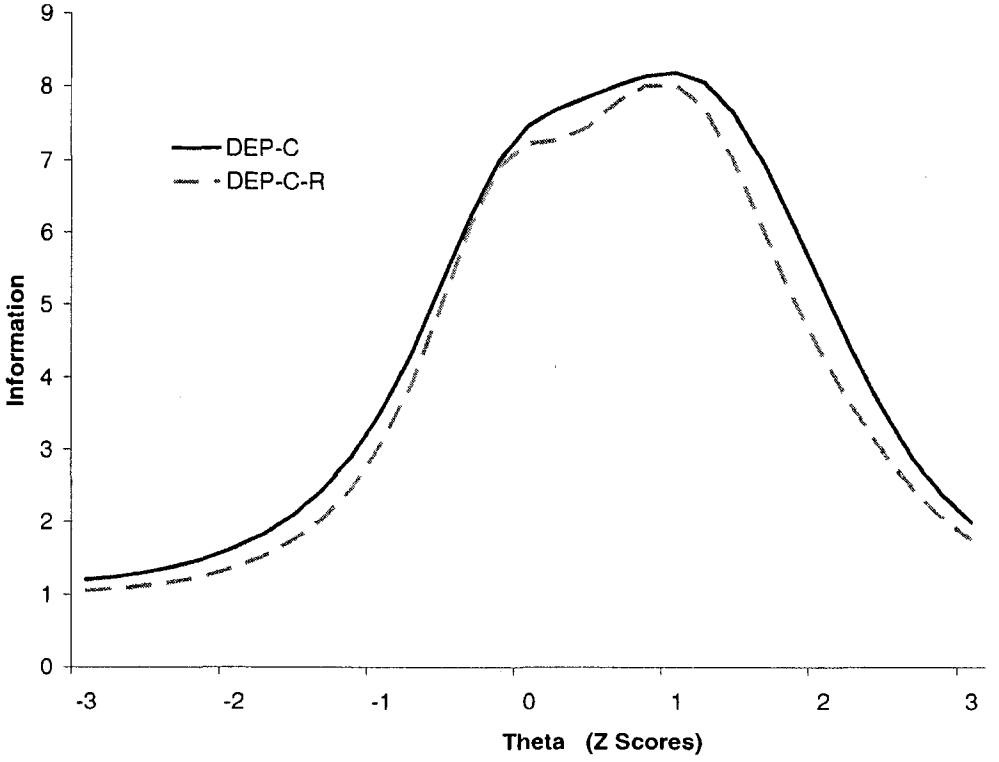
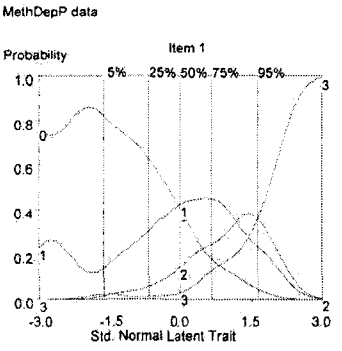
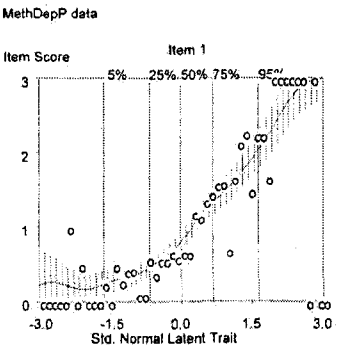


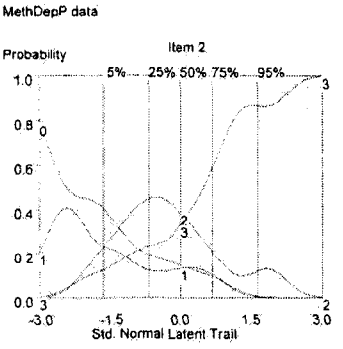
Figure 54. Comparison of total information curves for the eight item DEP-C as compared to the DEP-C-R based upon responses of a sample of MMT population (N = 323).

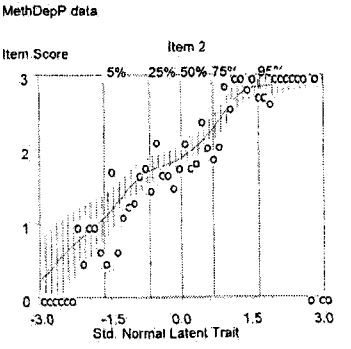
PAI Item #35



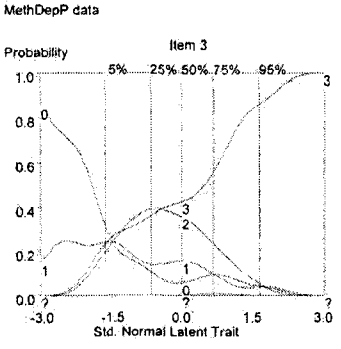


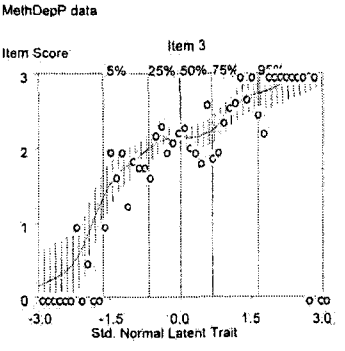
PAI Item #75



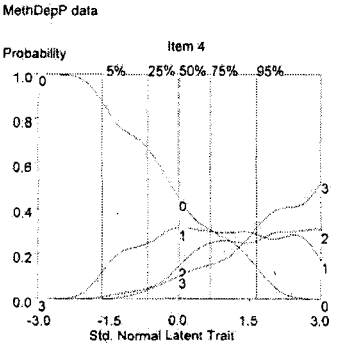


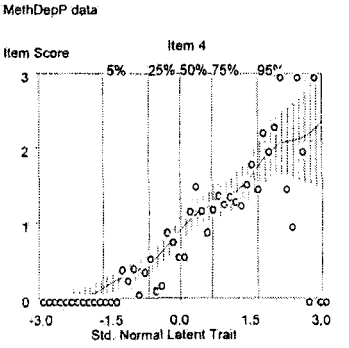
PAI Item #115



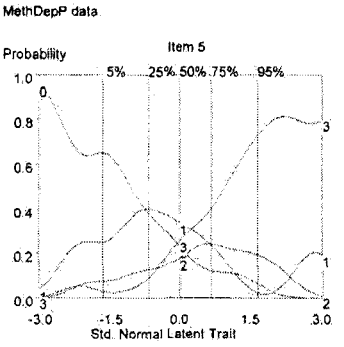


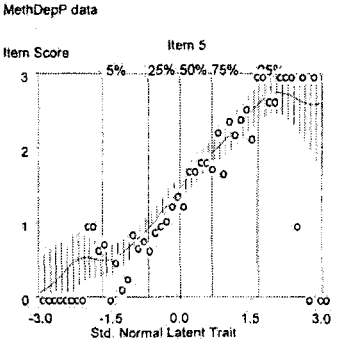
PAI Item #155



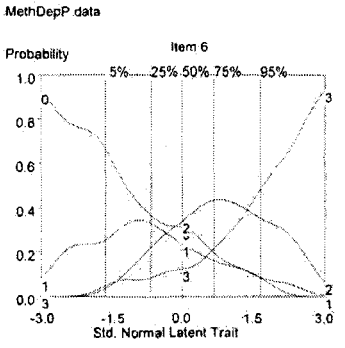


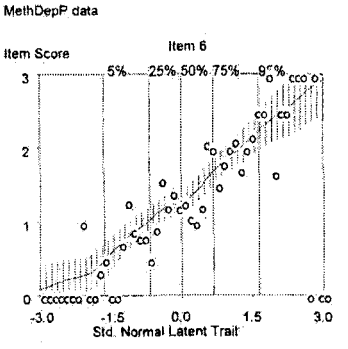
PAI Item #195



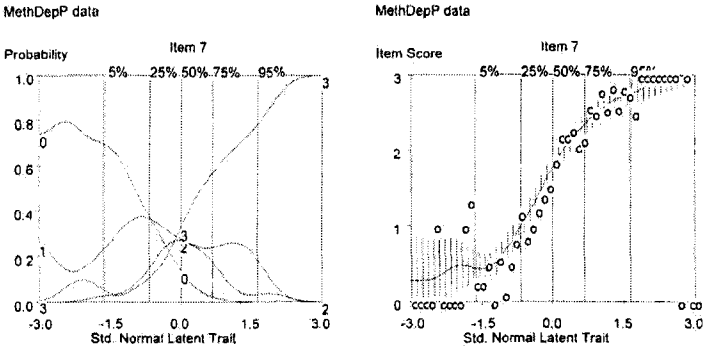


PAI Item #235





PAI Item #275



PAI Item #315

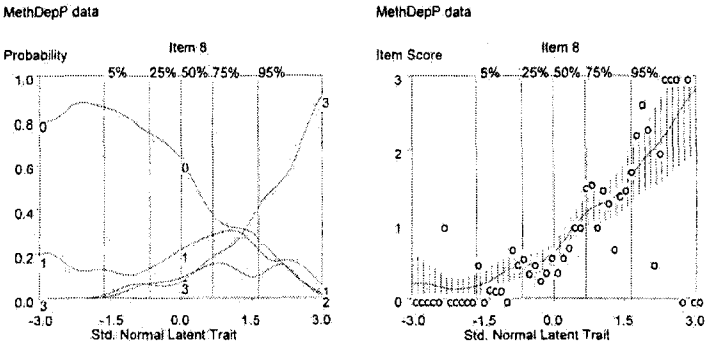


Figure 55. Option and item information curves for the DEP-P subscale plotted by Testgraf employing Gaussian kernel smoothing techniques based upon the responses of a sample of MMT patients ($N = 323$).

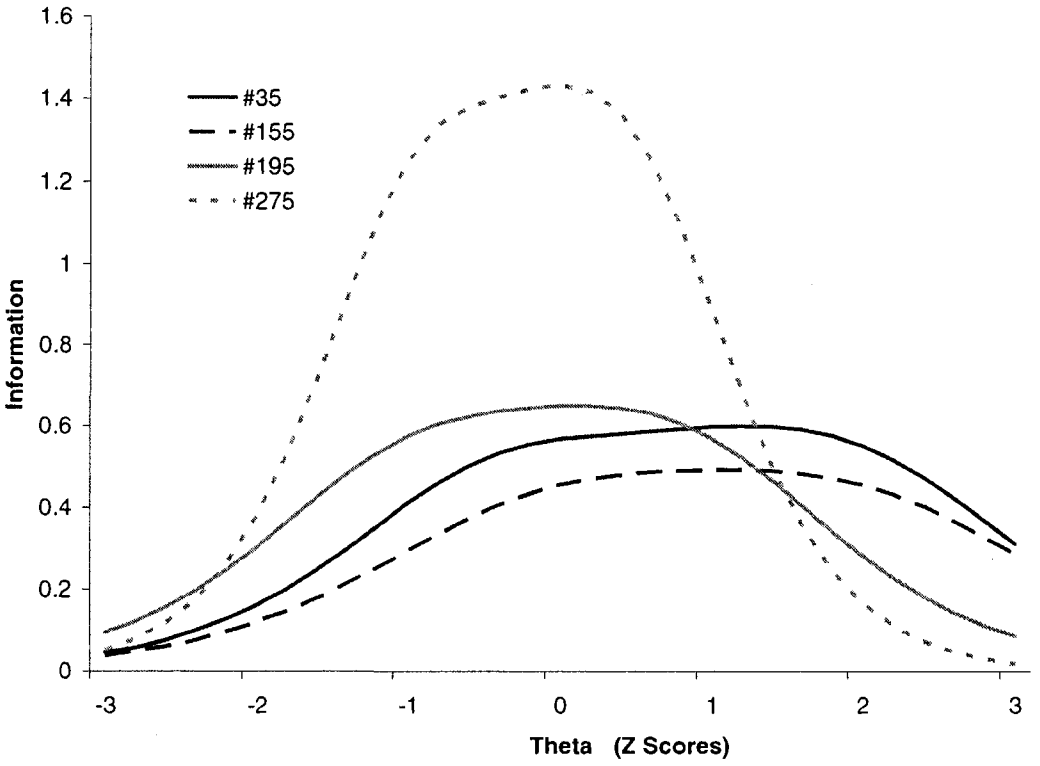


Figure 56. Item information curves for the DEP-P-R subscale based upon responses to the PAI in a sample of MMT patients (N = 323).

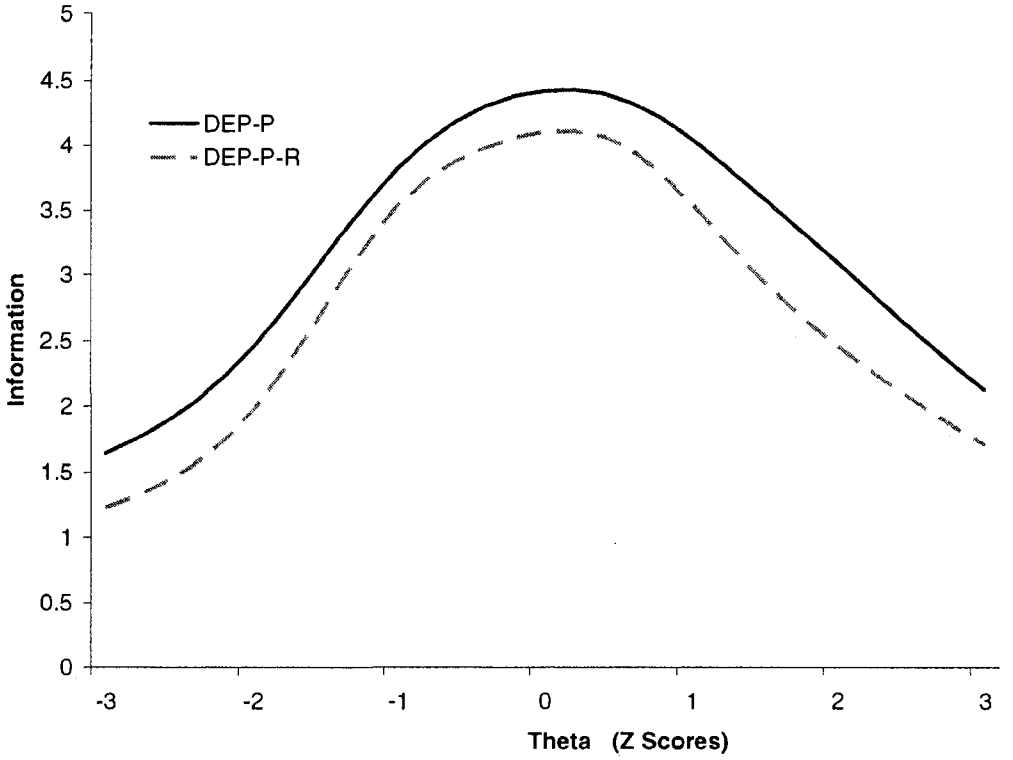
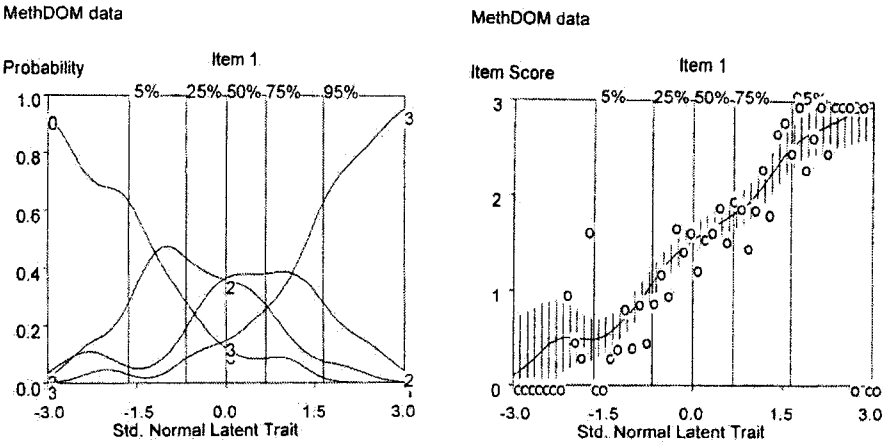
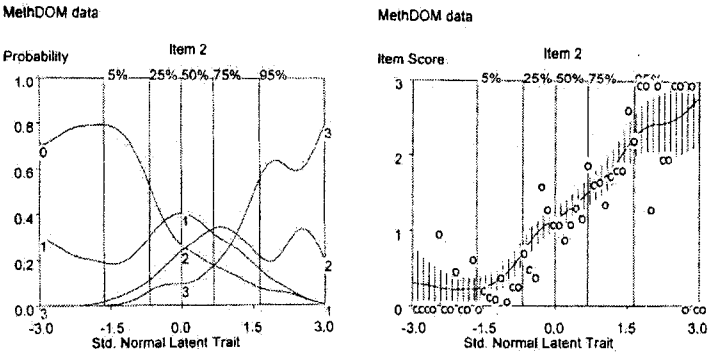


Figure 57. Total subscale information curves comparing the original eight item DEP-P and the new four item DEP-P-R subscale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

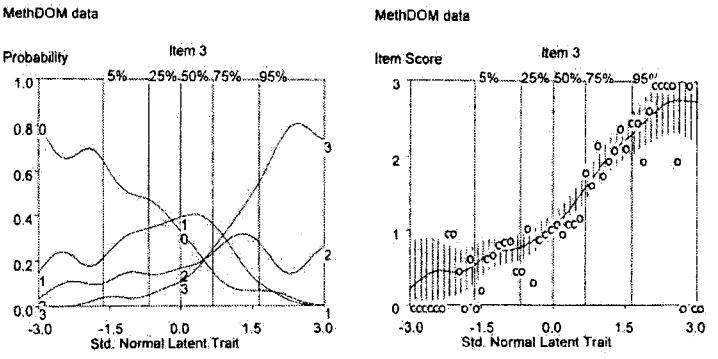
PAI Item #16



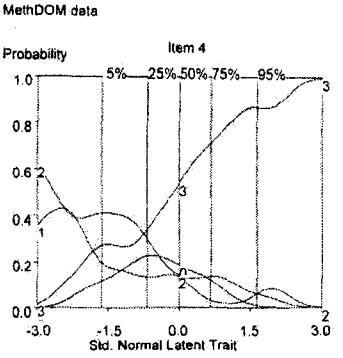
A. PAI Item #56

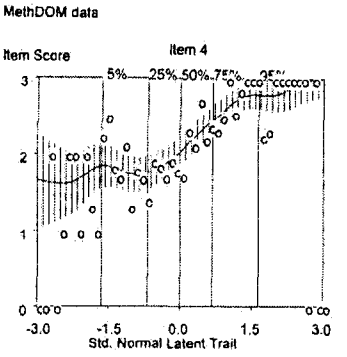


B. PAI Item #96

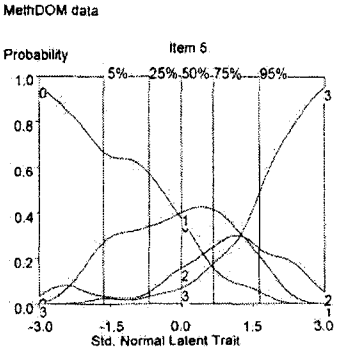


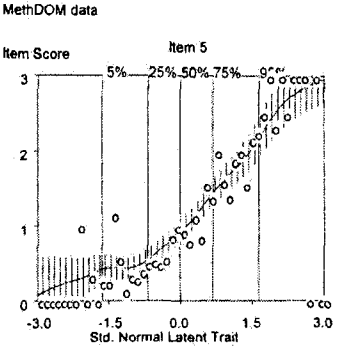
C. PAI Item #136



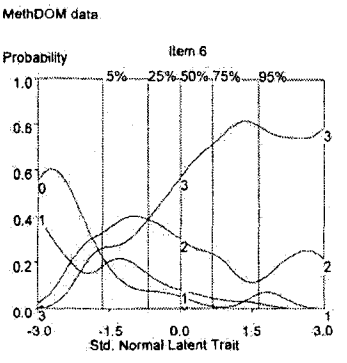


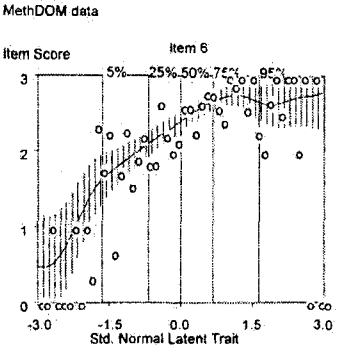
D. PAI Item #176



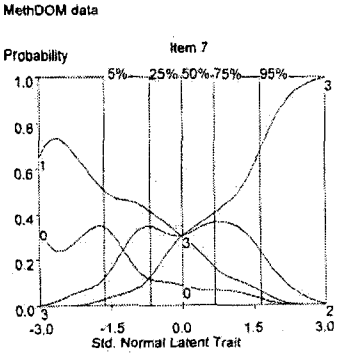


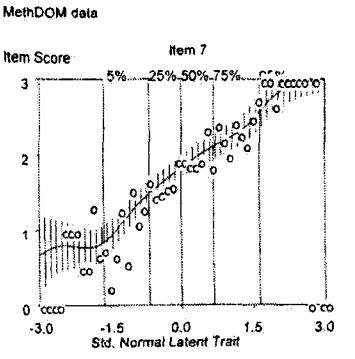
E. PAI Item #216



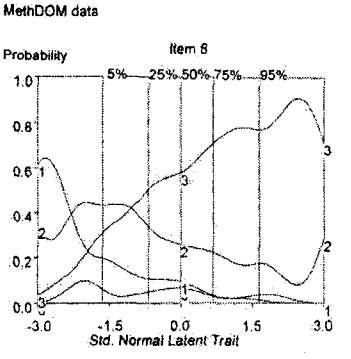


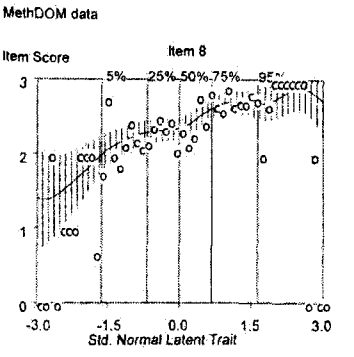
F. PAI Item #256



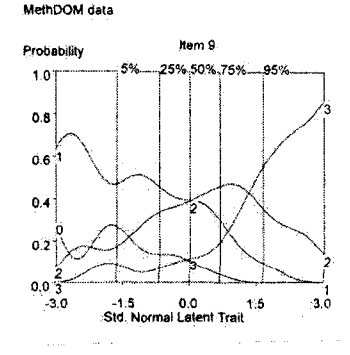


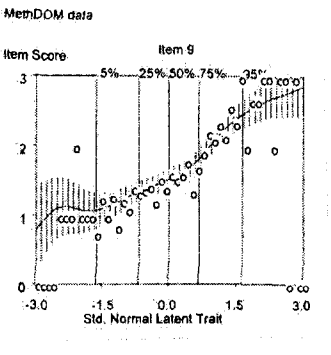
G. PAI Item #257



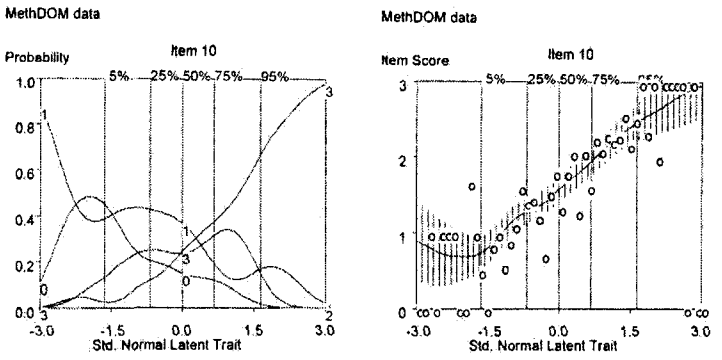


H. PAI Item # 296

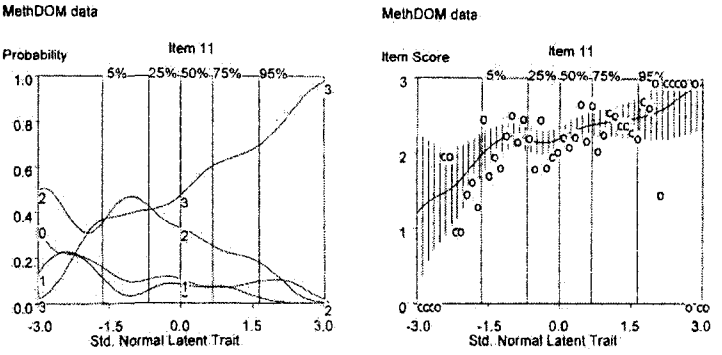




I. PAI Item #297



K. PAI Item #336



L. PAI Item #337

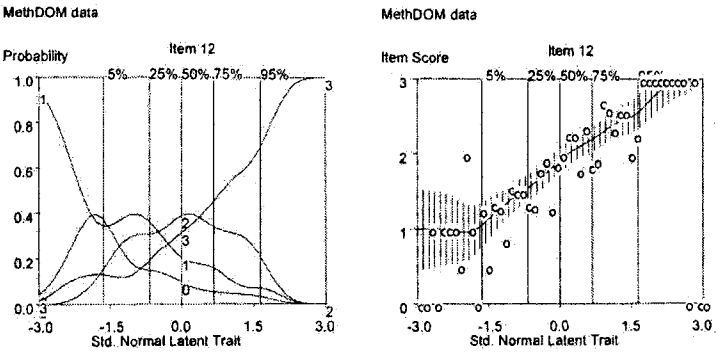


Figure 58. Option and item characteristic curves of the PAI scale DOM as plotted by Testgraf. Testgraf employs Gaussian kernel smoothing techniques to plot the data in a sample of MMT patients (N = 323).

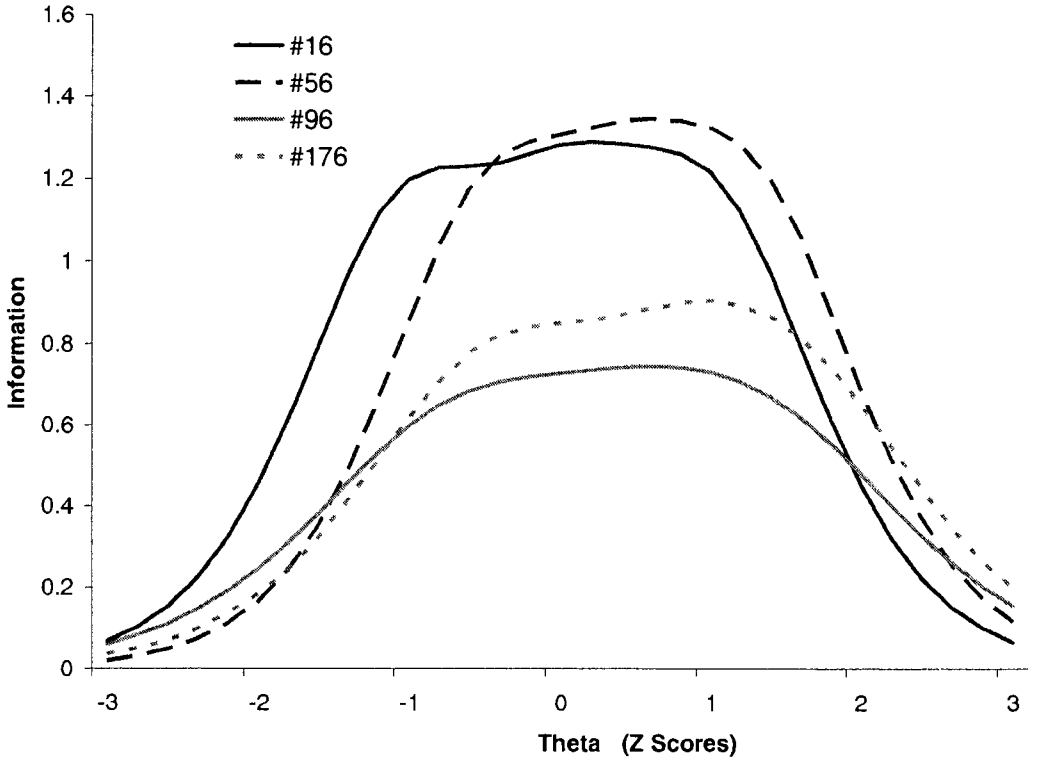


Figure 59. Item information curves for the revised DOM-R scale based upon responses to the PAI of a sample of MMT patients (N = 323).

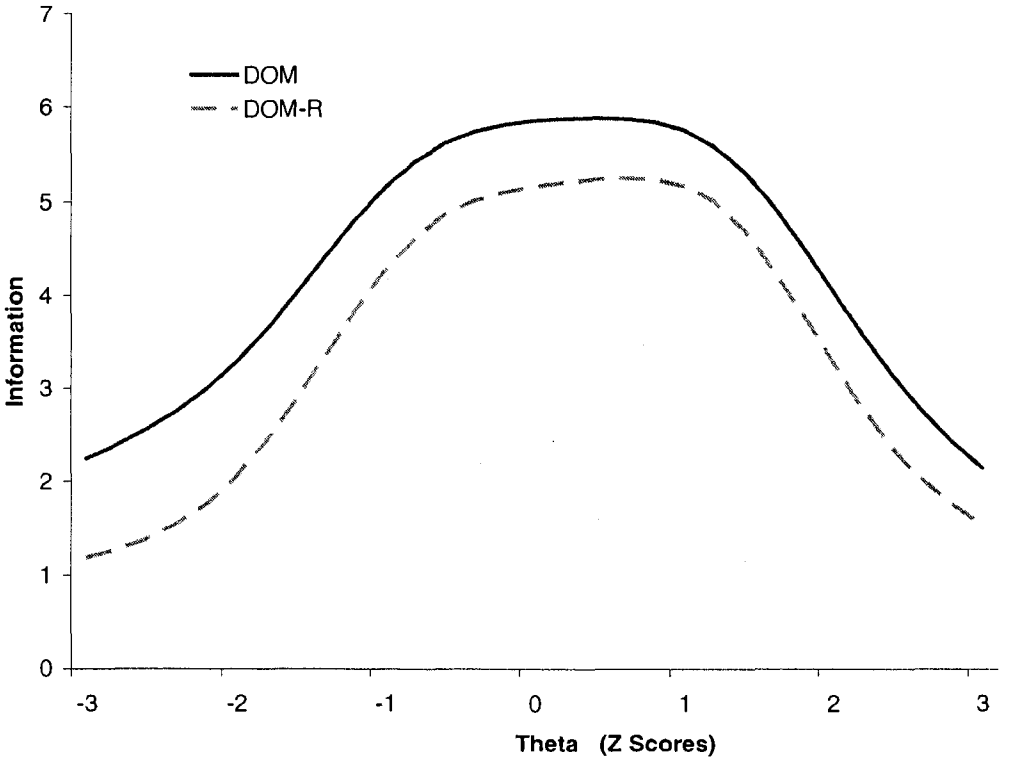
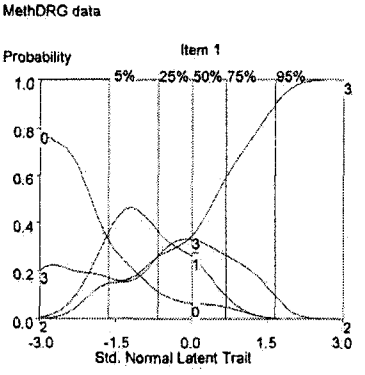
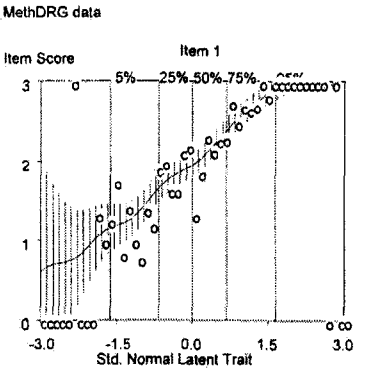


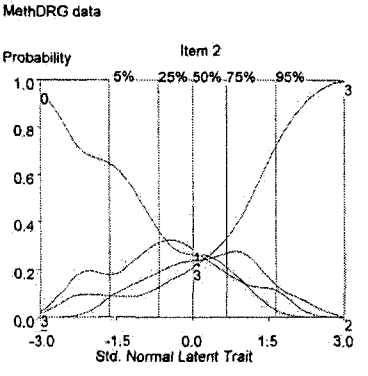
Figure 60. Comparison of estimated information provided by the original 12 item DOM scale and the newly revised DOM-R scale based upon responses to the PAI of a sample of MMT patients ($N = 323$).

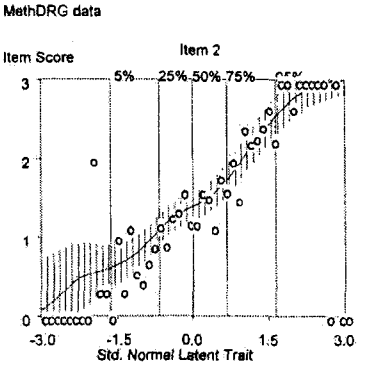
PAI Item #22



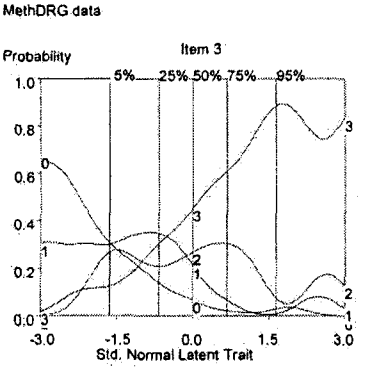


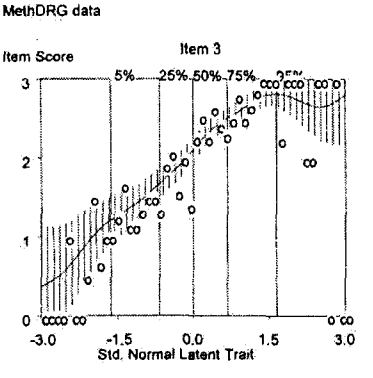
PAI Item #23



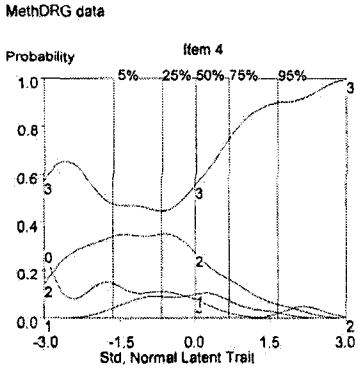


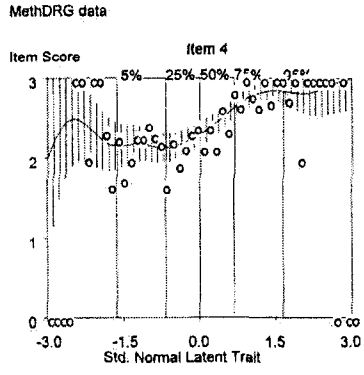
PAI Item #62



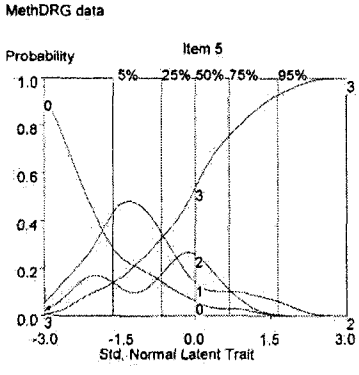


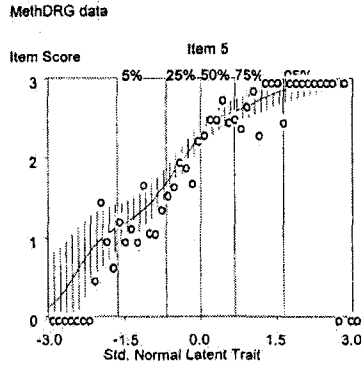
PAI Item #63



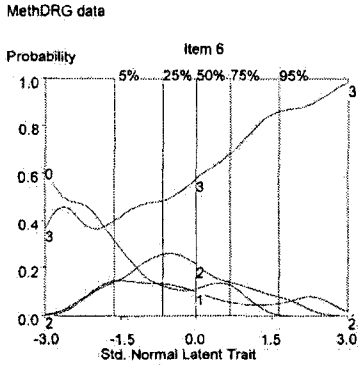


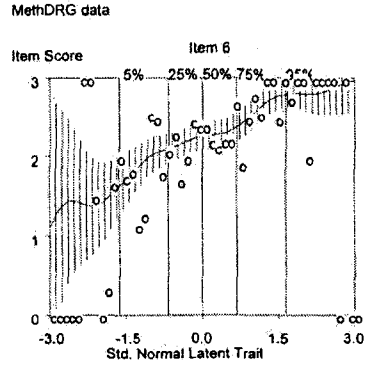
PAI Item #102



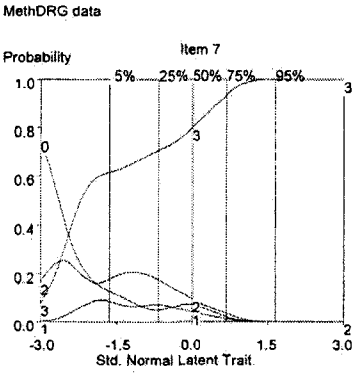


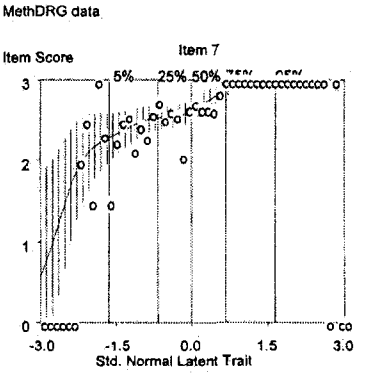
PAI Item #103



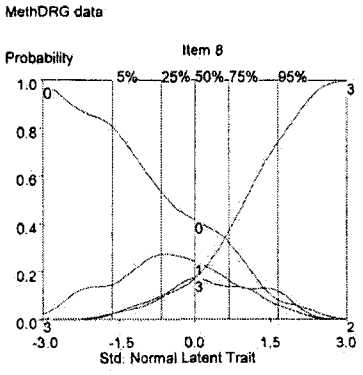


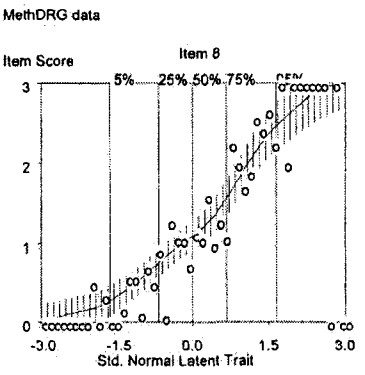
PAI Item #142



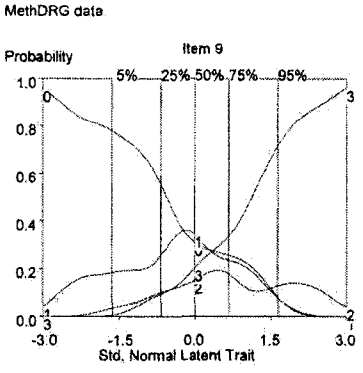


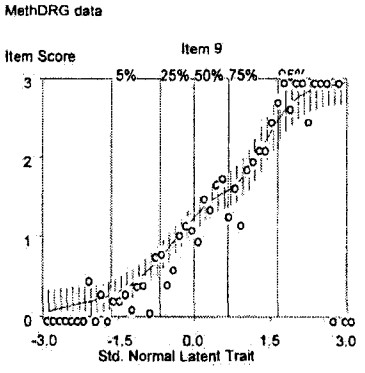
PAI Item #182



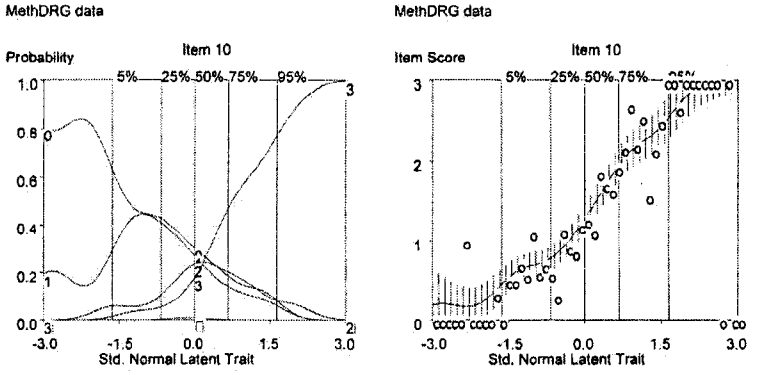


PAI Item #222

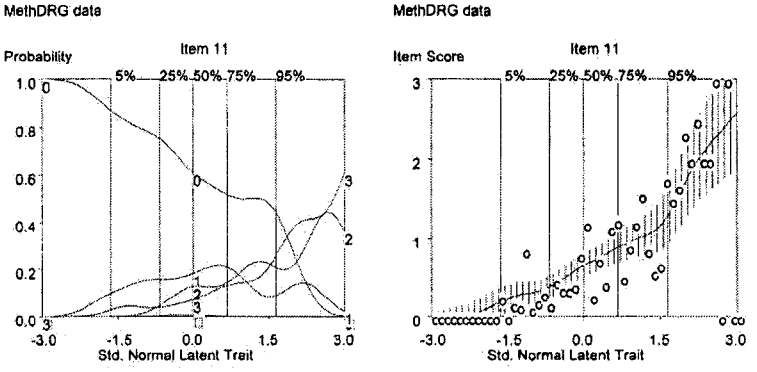




PAI Item #262



PAI Item #302



PAI Item #342

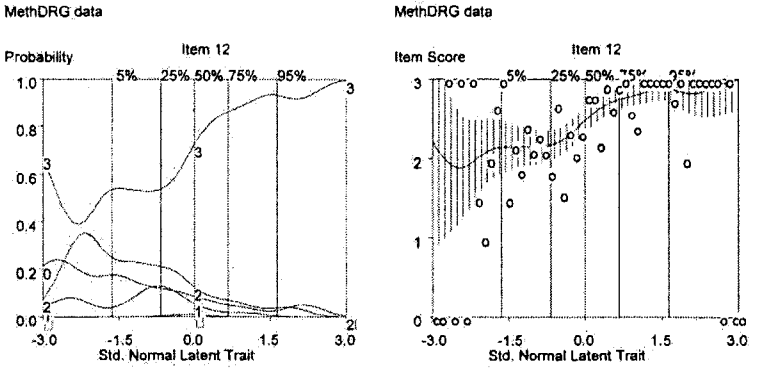


Figure 61. Option and item characteristic curves as plotted by the nonparametric IRT software Testgraf, employing Gaussian kernel smoothing techniques for the DRG scale of the PAI in a sample of MMT patients (N = 323).

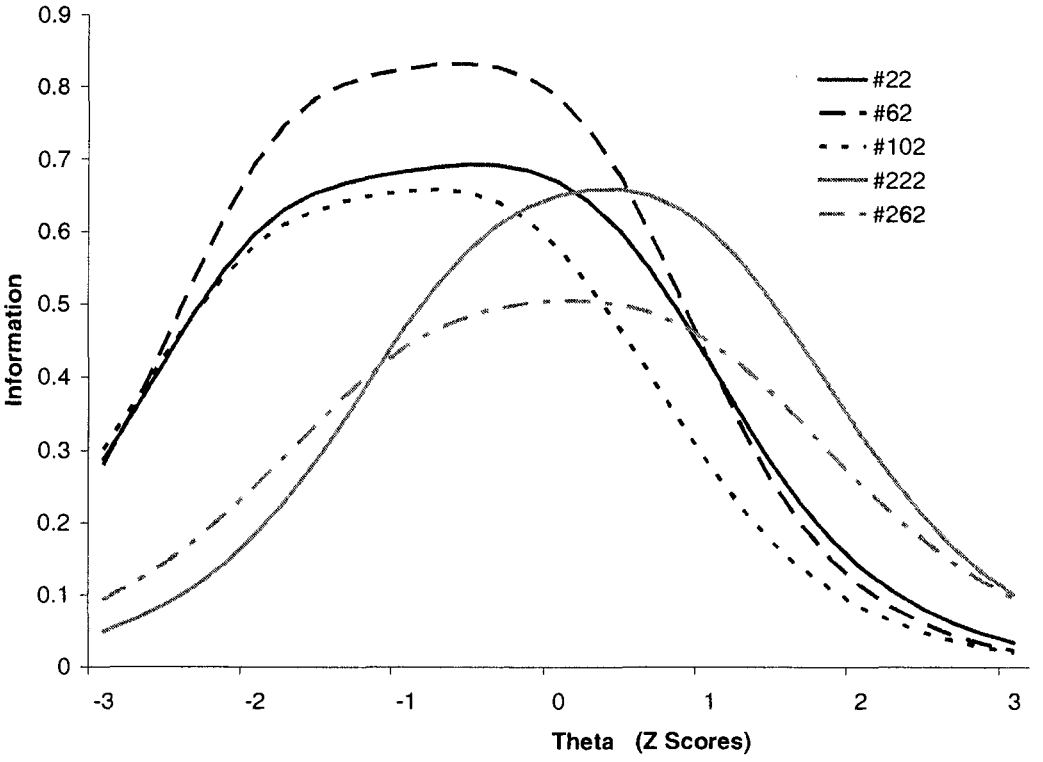


Figure 62. Item information curves for the DRG-R subscale based upon responses to the PAI in a sample of MMT patients ($N = 323$).

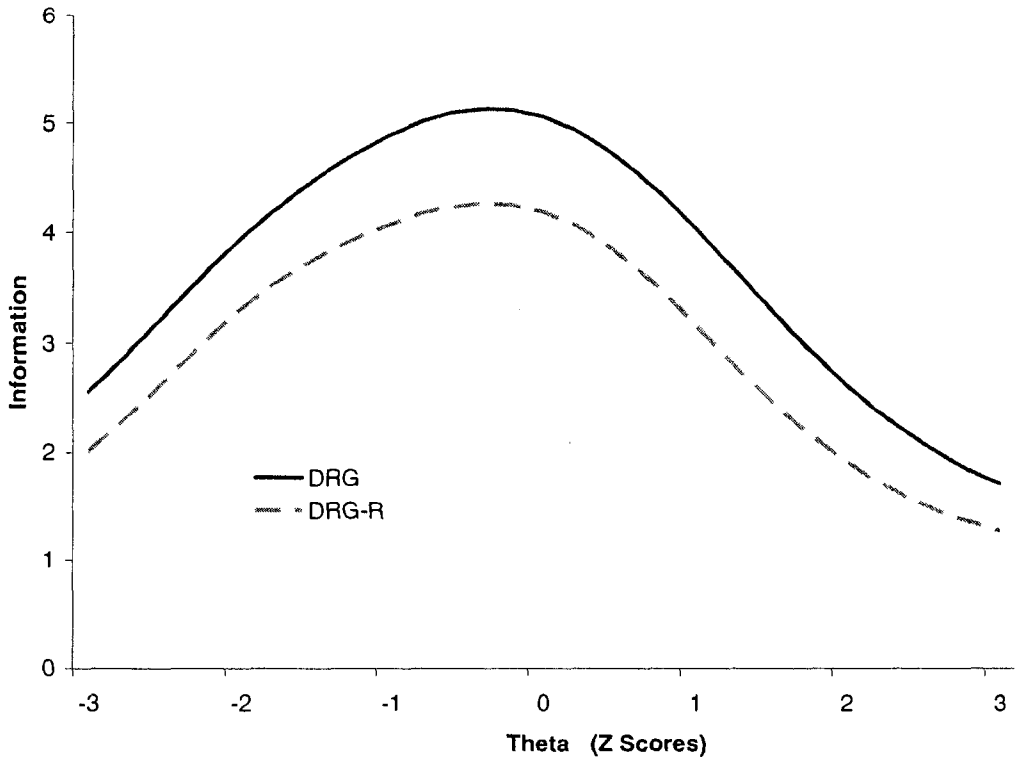
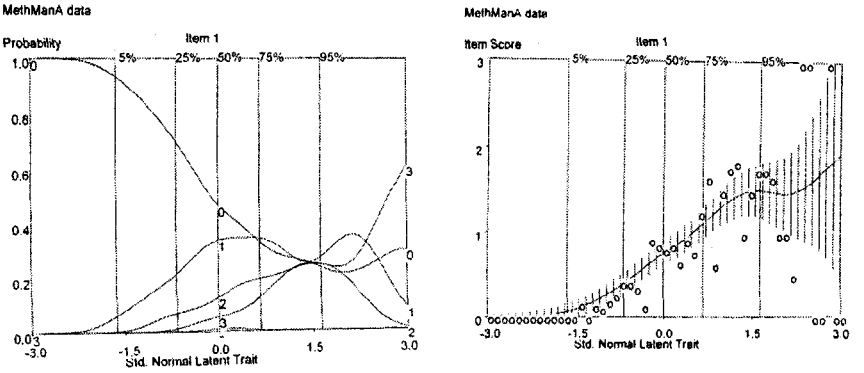
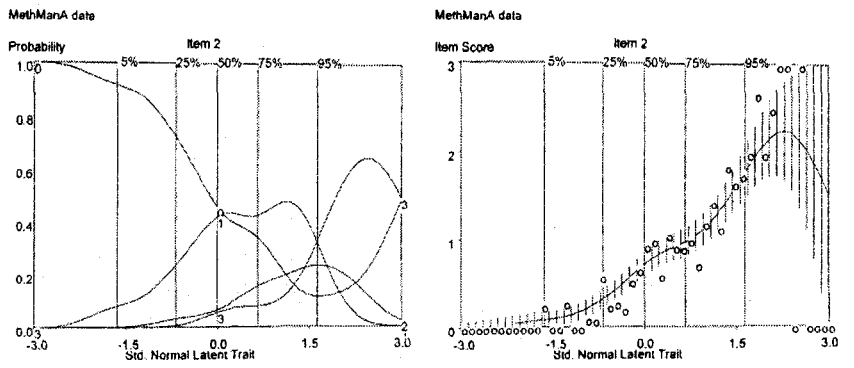


Figure 63. Subscale information curves for the newly derived DRG-R (5 items) scale as compared to the original 12 item scale based upon responses to the PAI in a sample of MMT patients ($N = 323$).

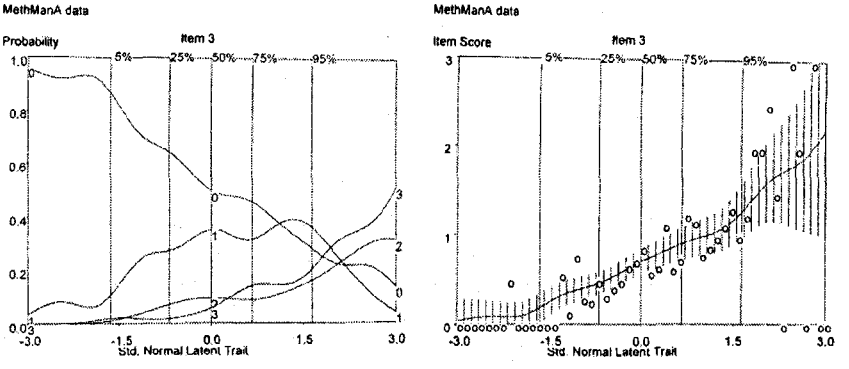
A. Item #7



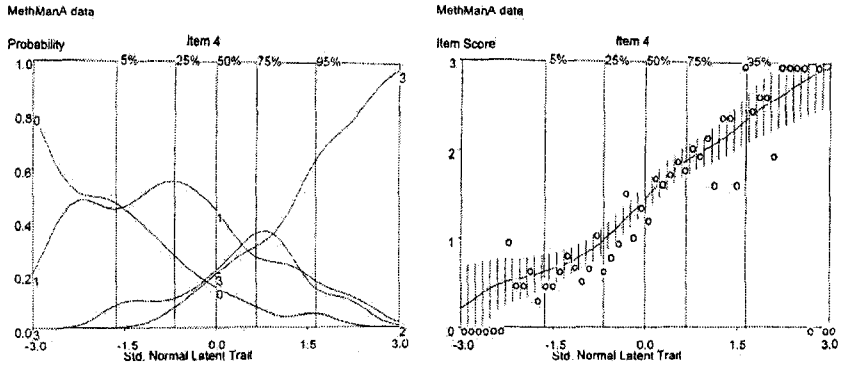
B. Item #47



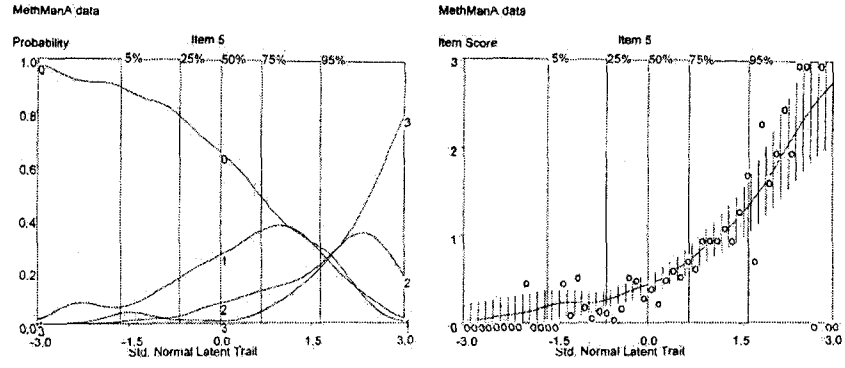
C. Item #87



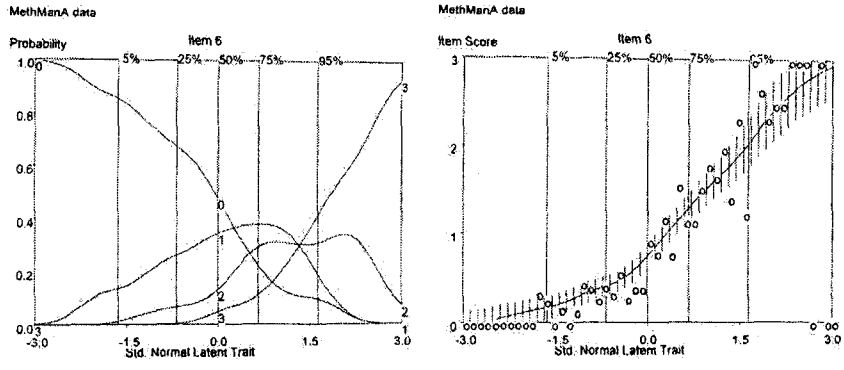
D. Item #127



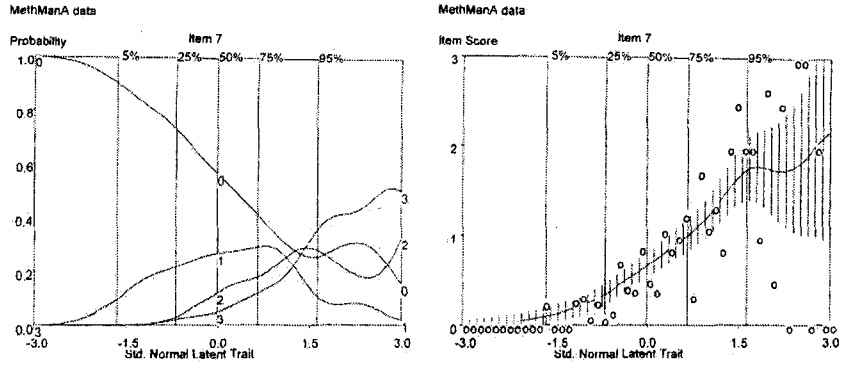
E. Item #167



F. Item #207



G. Item #247



H. Item #287

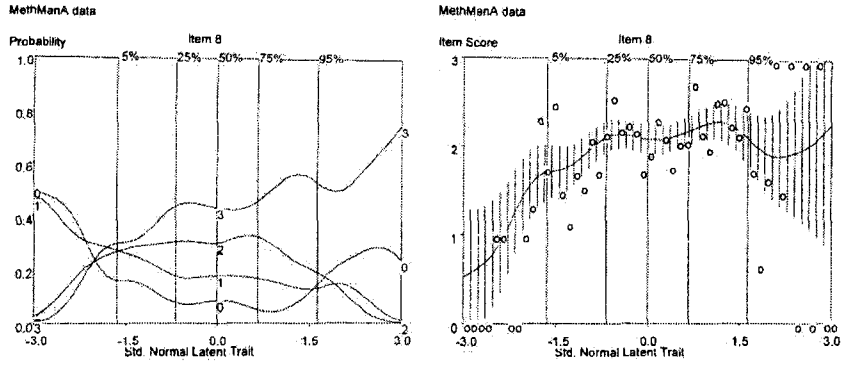


Figure 64. Option and item characteristic curves for the items of the MAN-A subscale as plotted by Testgraf employing Gaussian kernel smoothing techniques in a sample of MMT patients (N = 323).

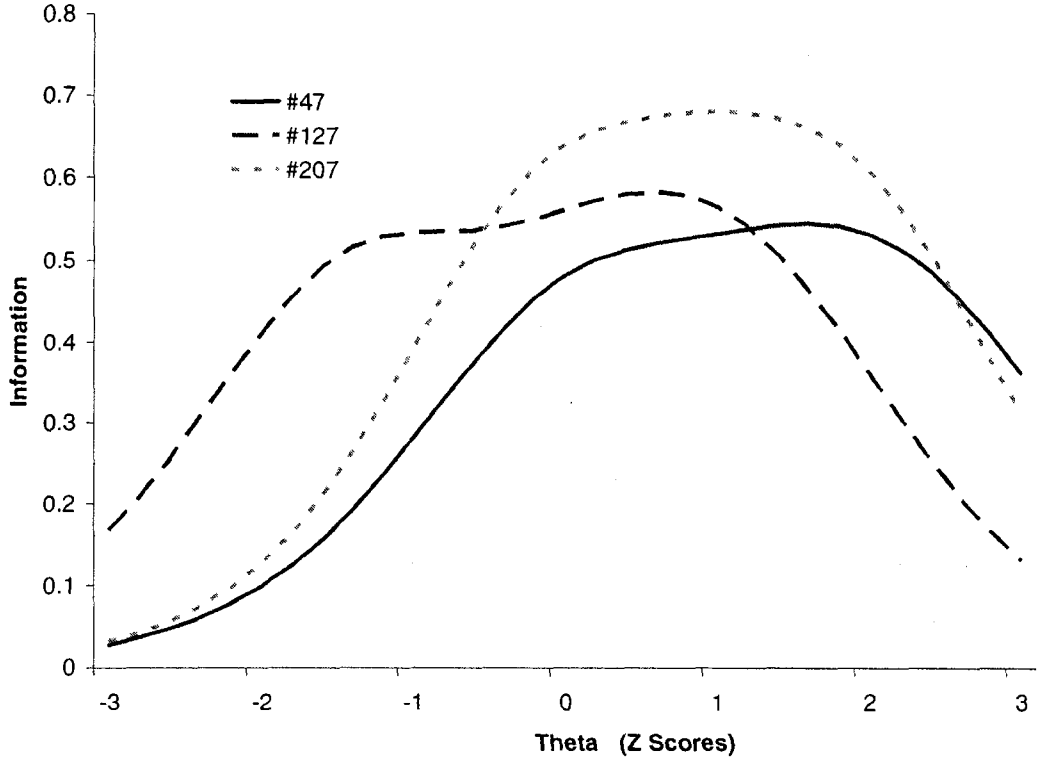


Figure 65. Item information curves for the three items of the MAN-A subscale based upon responses to the PAI of a sample of MMT patients (N = 323).

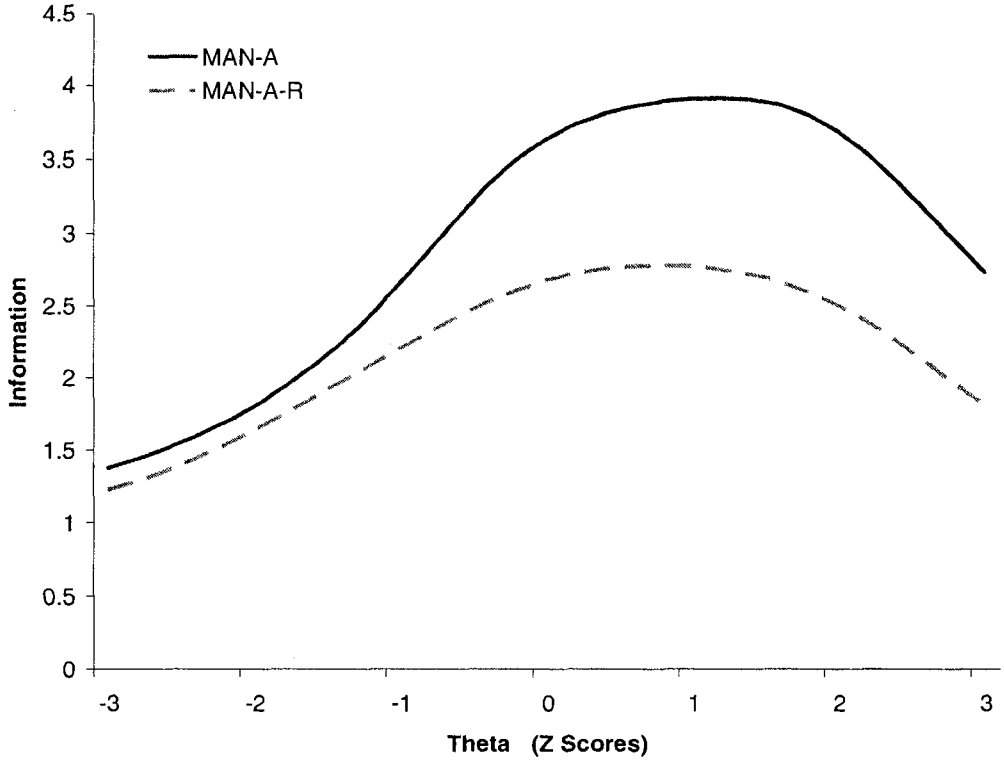
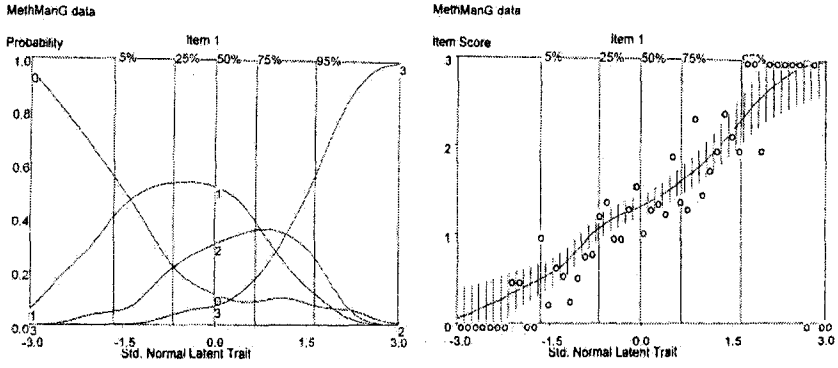
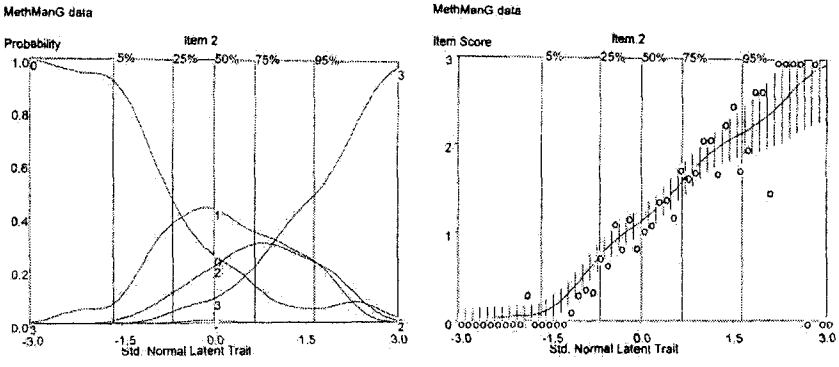


Figure 66. Total subscale information curves comparing the eight item MAN-A subscale and the three item MAN-A-R subscale based upon responses of a sample of MMT patients ($N = 323$).

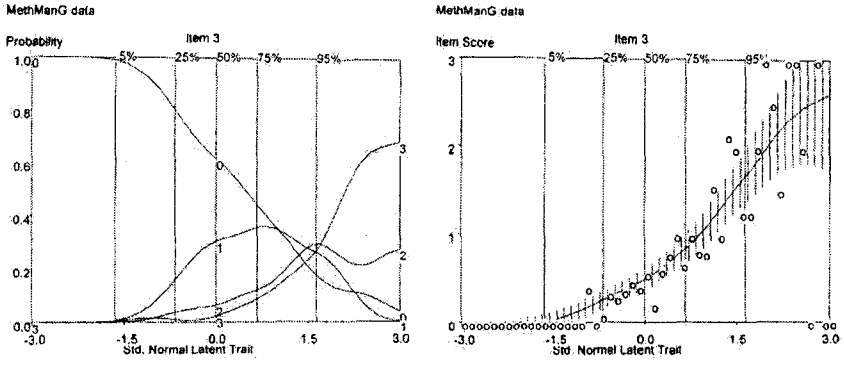
A. PAI Item #28



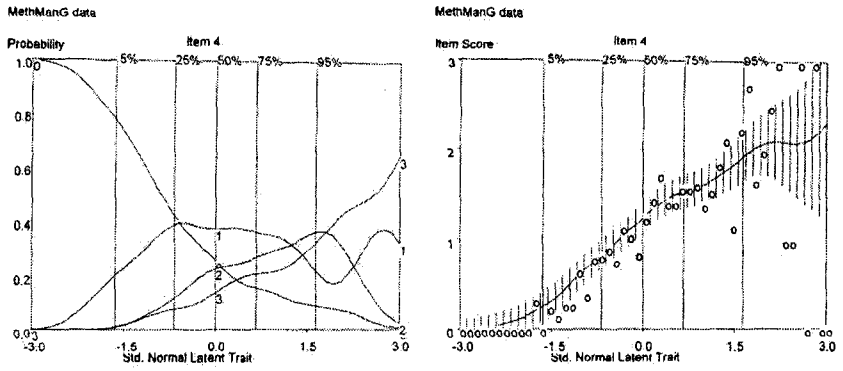
B. PAI Item #68



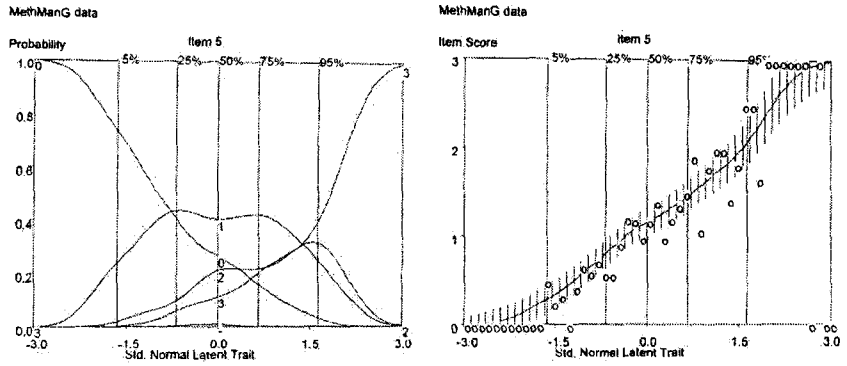
C. PAI Item #108



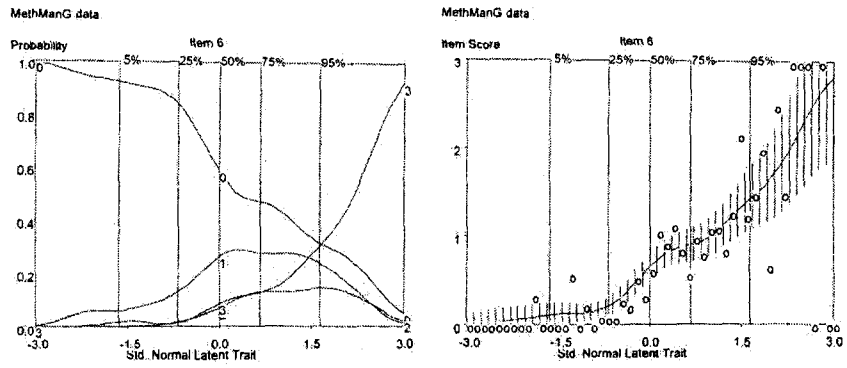
D. PAI Item #148



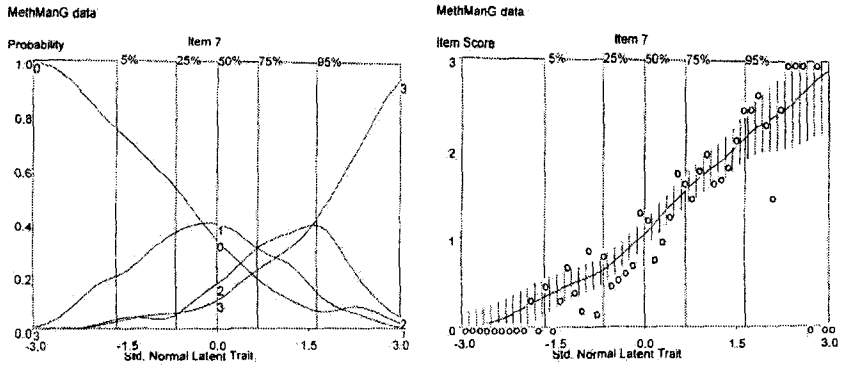
E. PAI Item #188



F. PAI Item #228



G. PAI Item #268



H. PAI Item #308

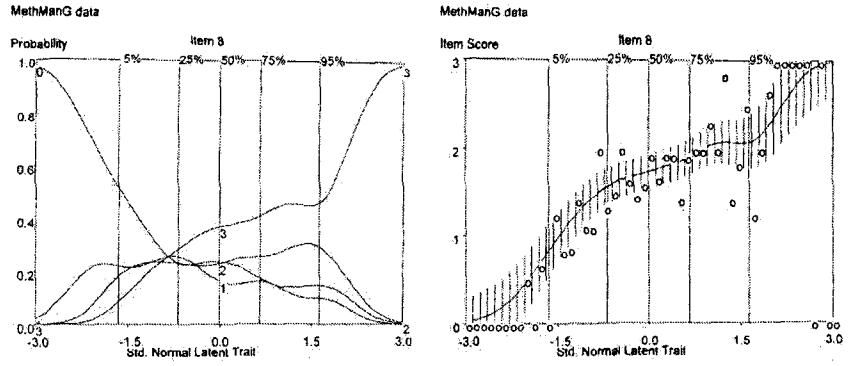


Figure 67. Option and item characteristic curves for the MAN-G subscale of the PAI as plotted by Testgraf employing a Gaussian Kernel smoothing techniques, based upon responses to the PAI by a sample of MMT patients ($N = 323$).

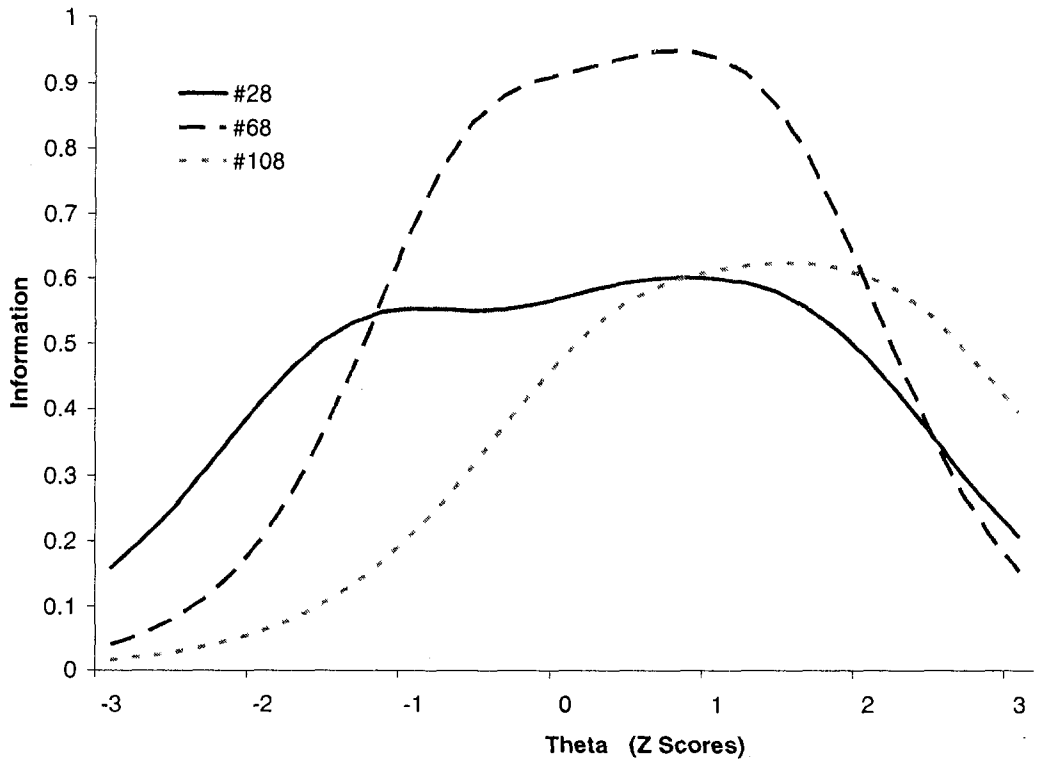


Figure 68. Item information curves of the revised MAN-G-R subscale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

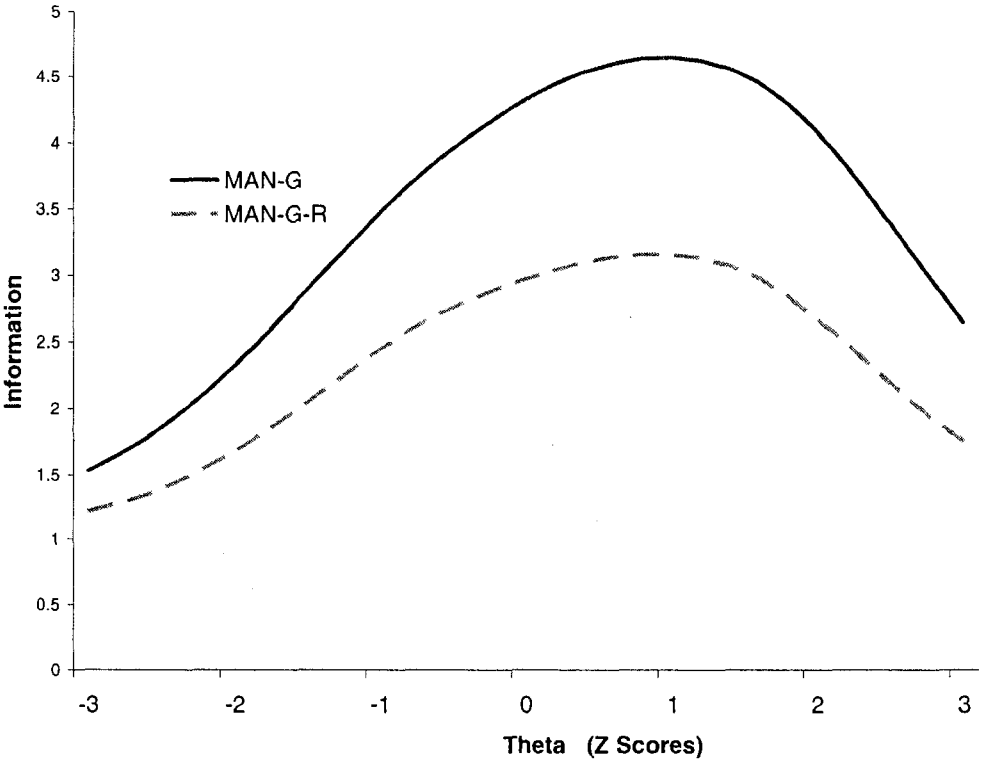
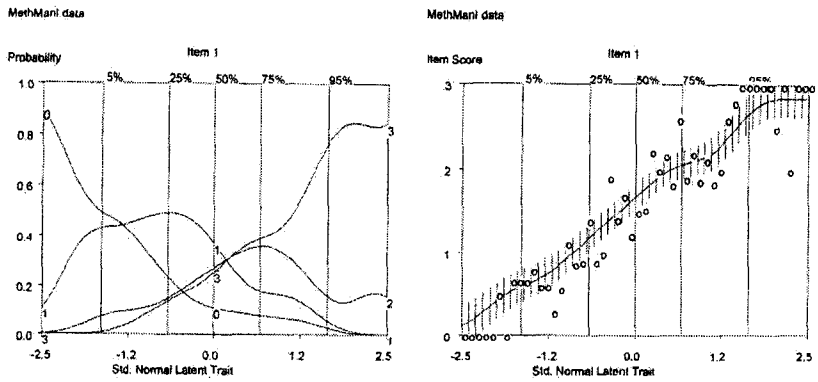
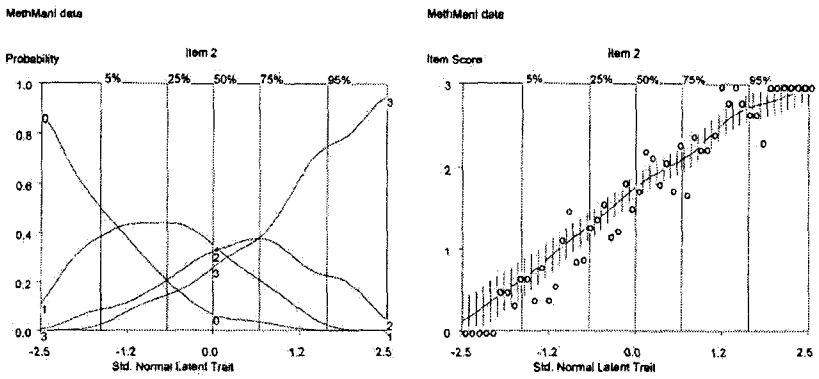


Figure 69. Total subscale information curves comparing the eight item MAN-G subscale to the revised three item MAN-G-R subscale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

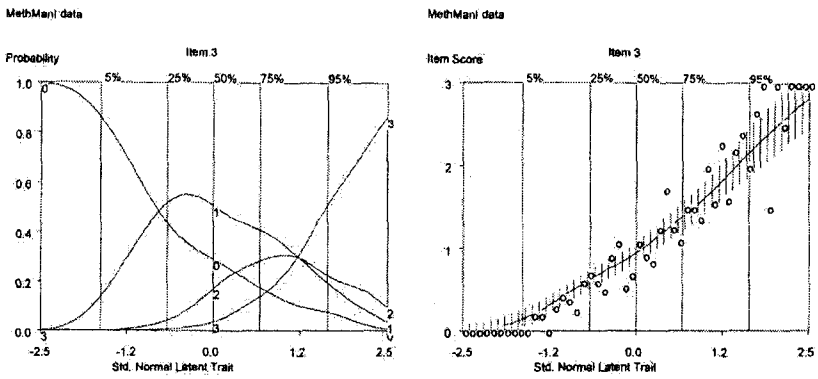
A. PAI Item #36



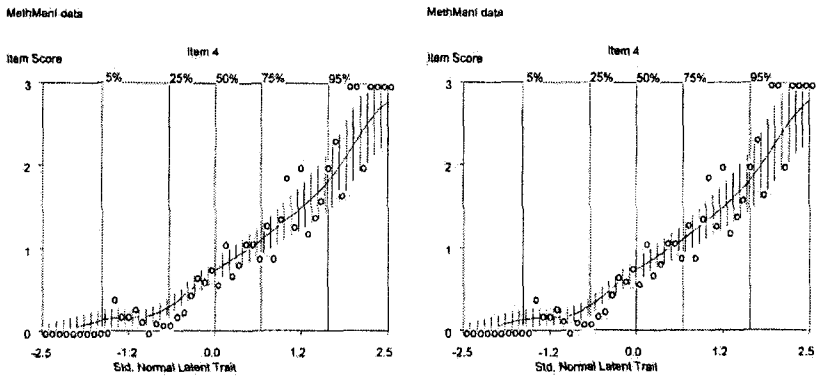
B. PAI Item #76



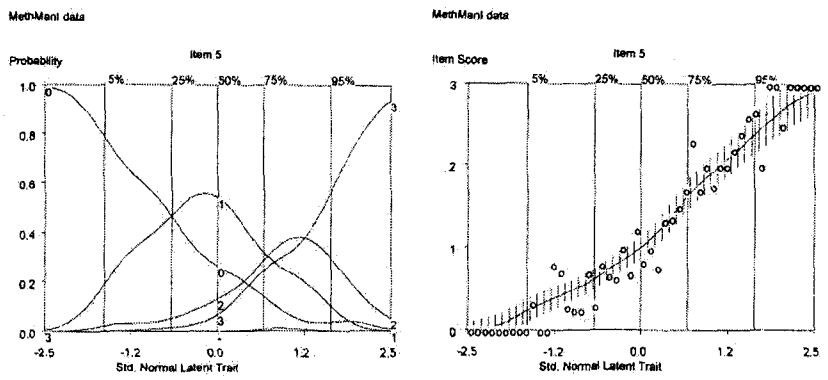
C. PAI Item #116



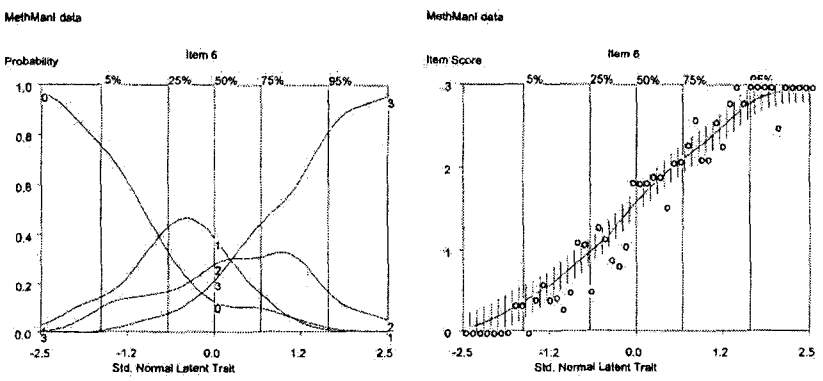
D. PAI Item #156



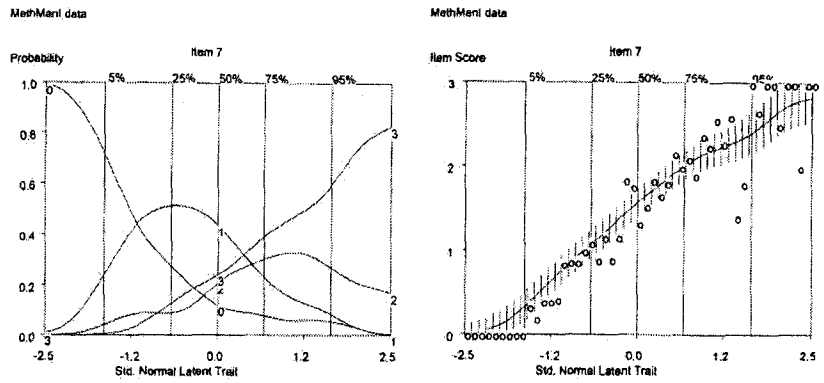
E. PAI Item #196



F. PAI Item #236



G. PAI Item #276



H. PAI Item #316

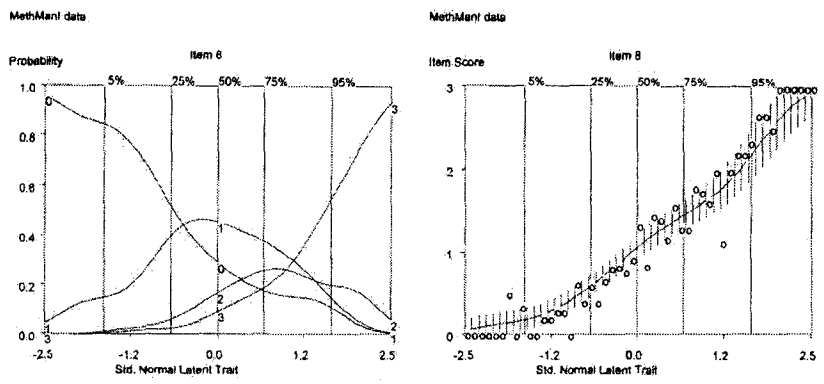


Figure 70. Option and item information curves for the items of the PAI subscale MAN-I as plotted by Testgraf employing a Gaussian Kernel smoothing technique based upon responses by a sample of MMT patients (N = 323).

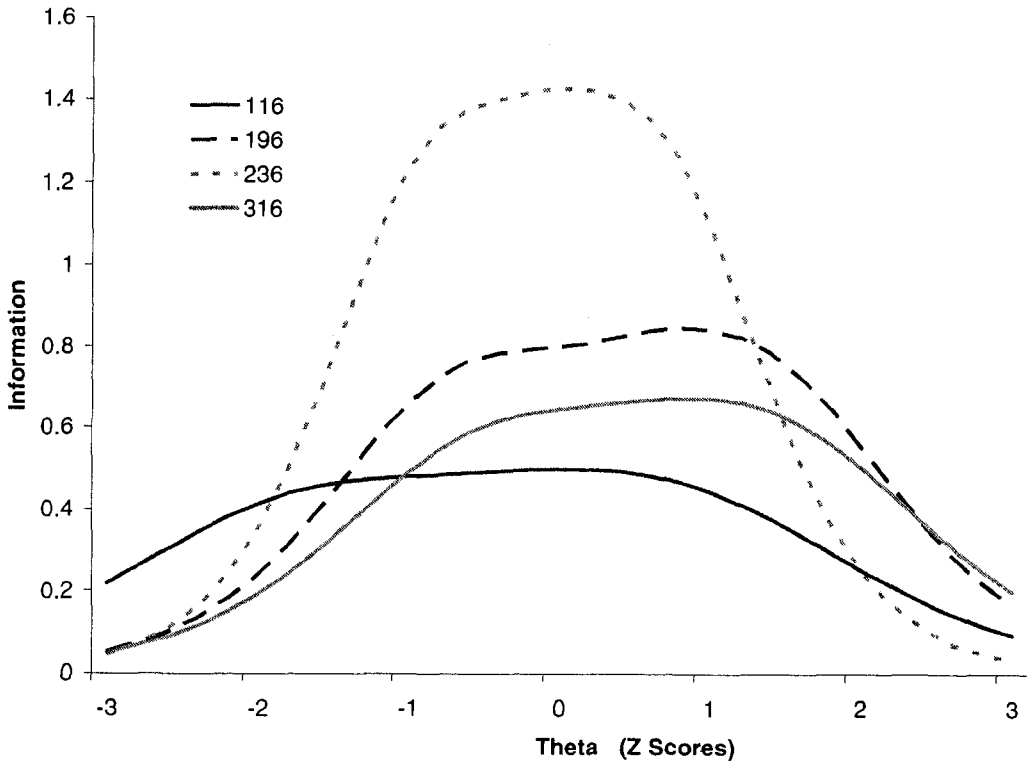


Figure 71. Item information curves for the new MAN-I-R subscale based upon responses to the PAI from a sample of MMT patients ($N = 323$).

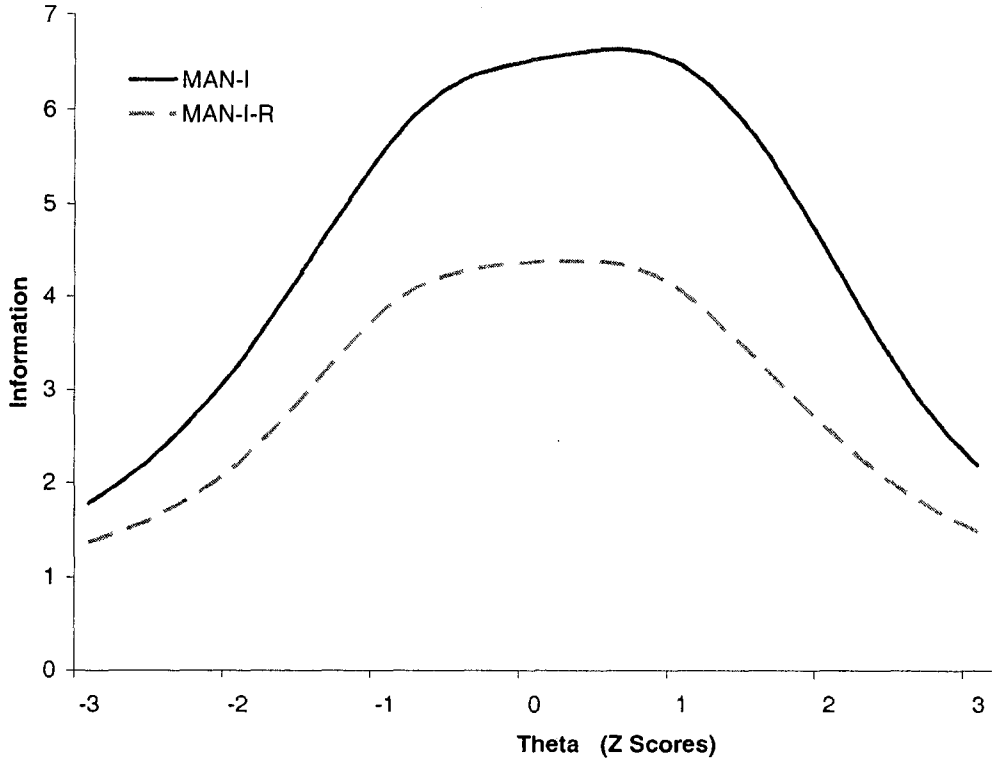
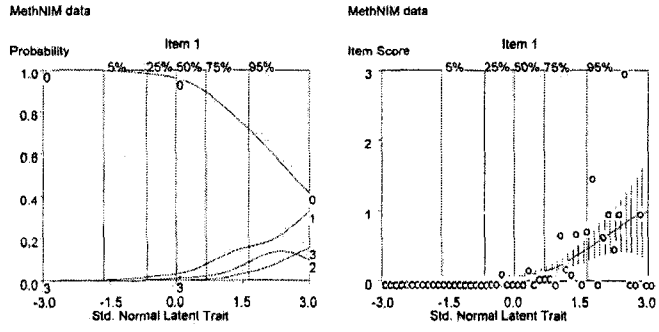
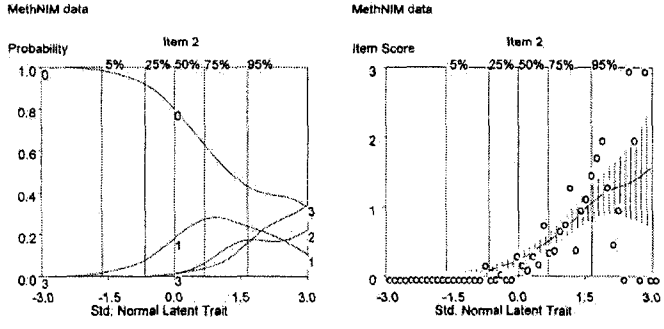


Figure 72. Total subscale information curves comparing the total information of the MAN-I subscale as compared to the MAN-I-R subscale based upon responses to the PAI by a MMT patient sample ($N = 323$).

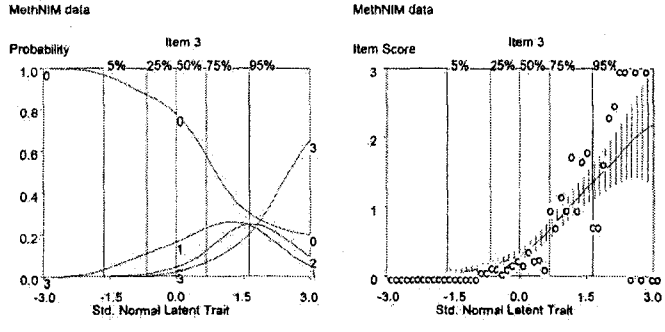
PAI Item # 9



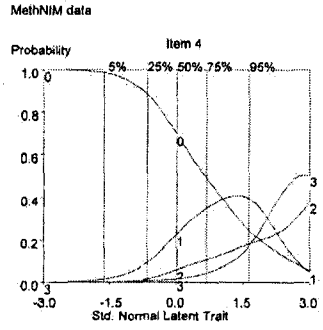
PAI Item # 49

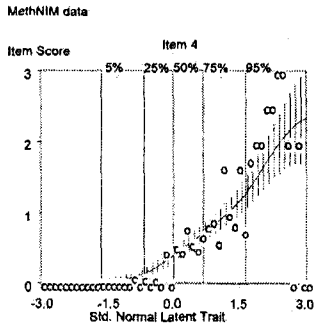


PAI Item # 89

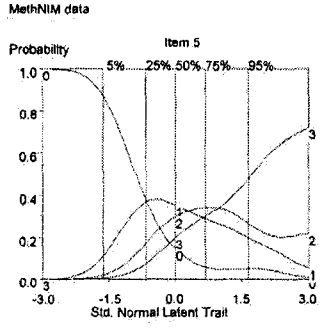


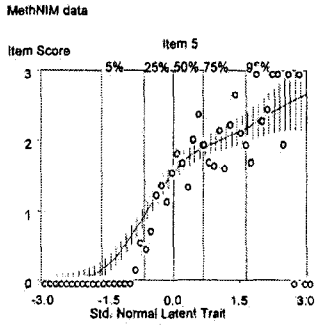
PAI Item # 129



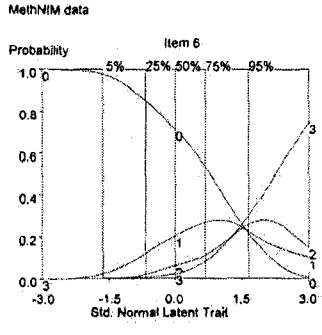


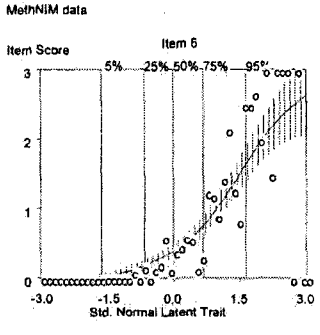
PAI Item # 169



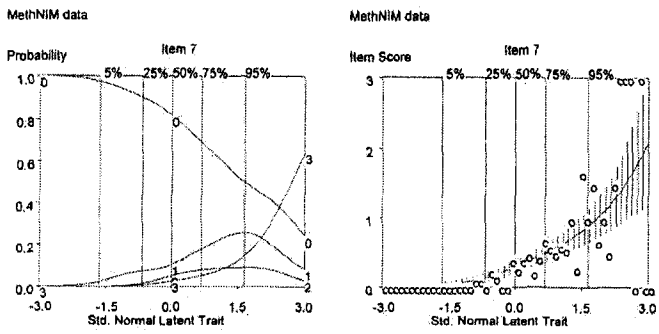


PAI Item # 209

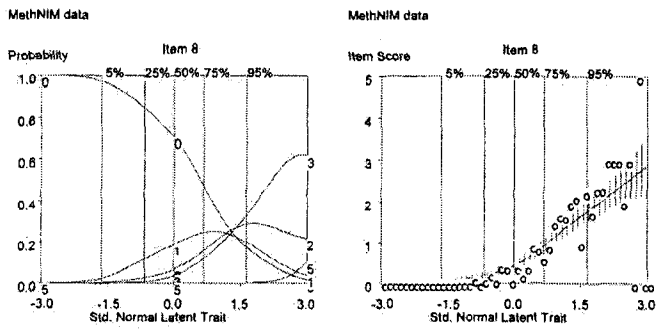




PAI Item # 249



PAI Item # 289



PAI Item # 329

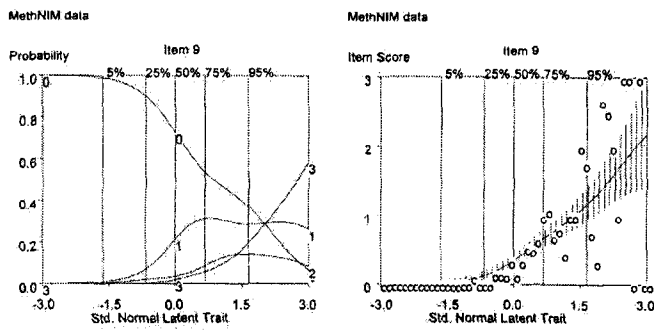


Figure 73. Option characteristic curves (OCC) and Item characteristic Curves (ICC) for the each item of the NIM scale of the PAI for the methadone maintenance treatment population (N = 323).

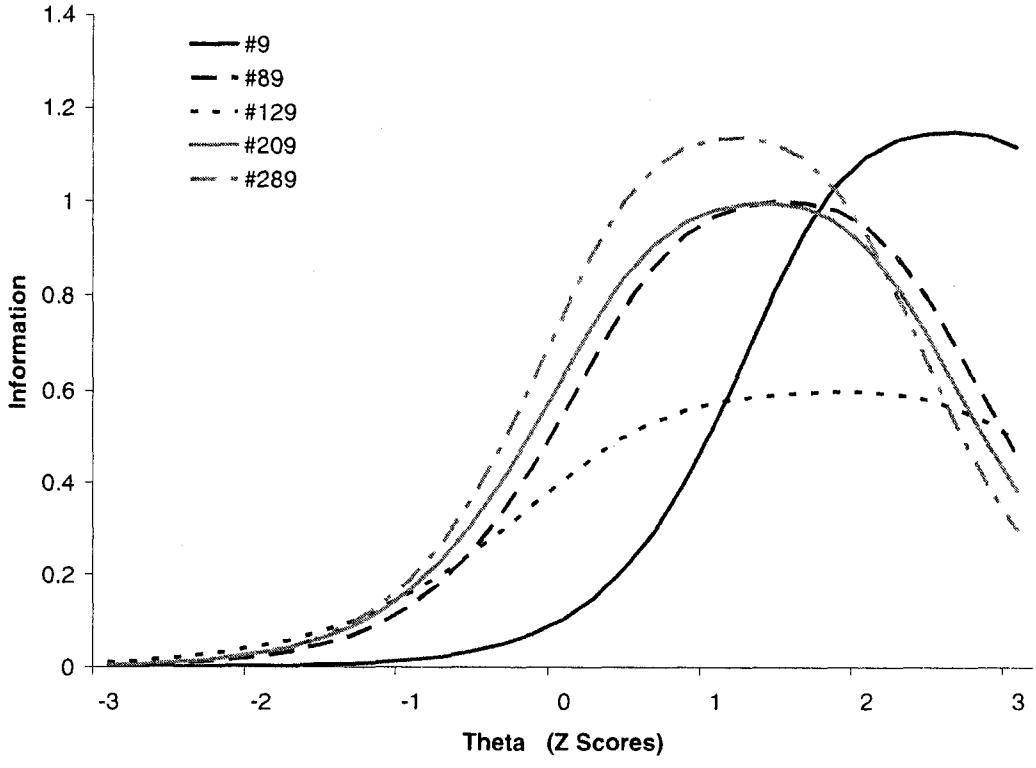


Figure 74. Item information curves for the final set of items for the NIM scale in use with a MMT population (N = 323).

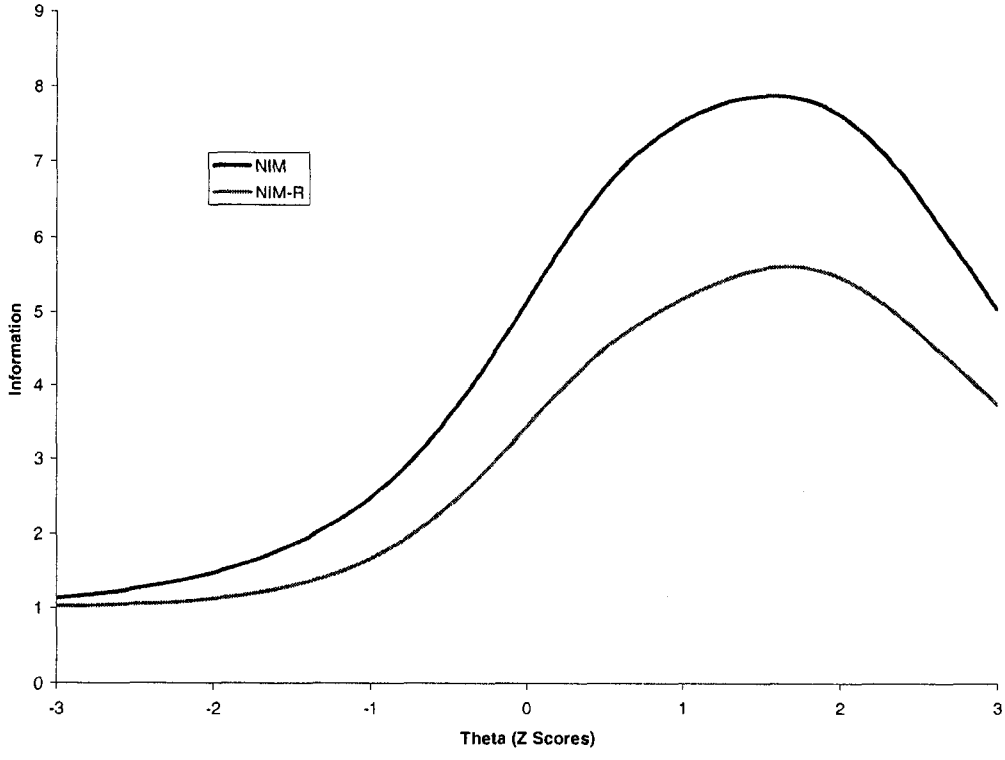
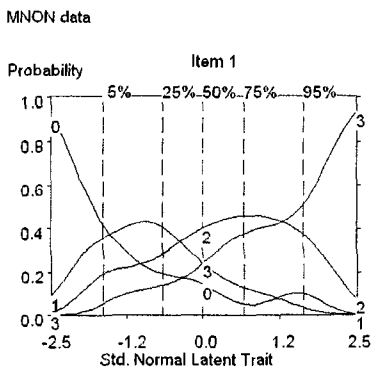
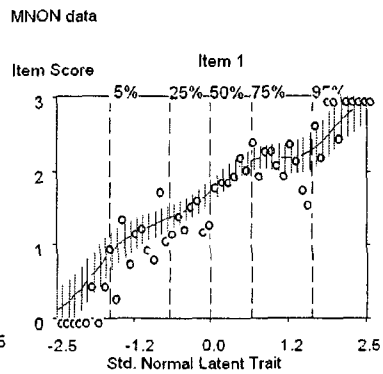


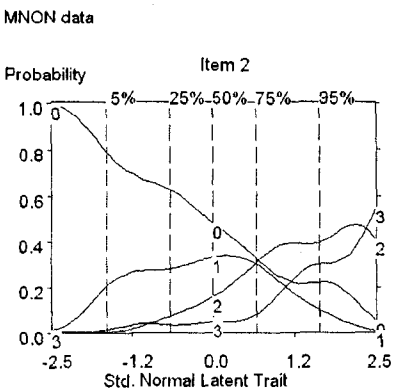
Figure 75. Total information curves comparing the total information offered by the original nine item NIM scale and the five item NIM-R based upon responses to the PAI by a sample of MMT patients. (N = 323).

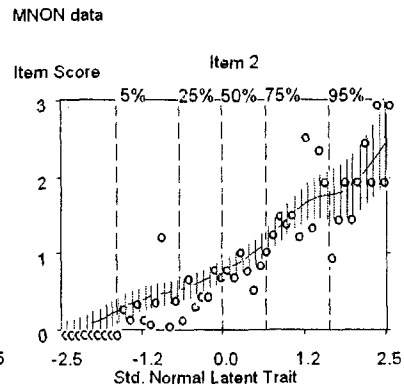
PAI Item #1



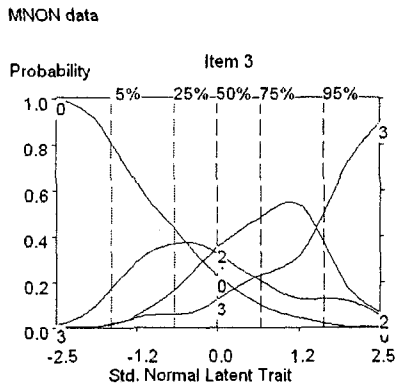


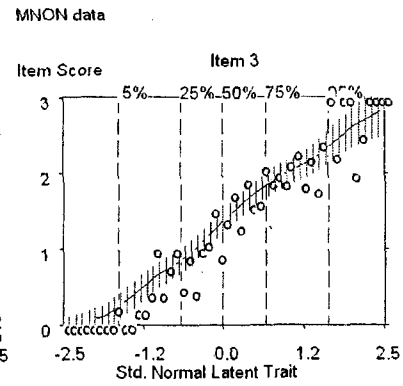
PAI Item #41



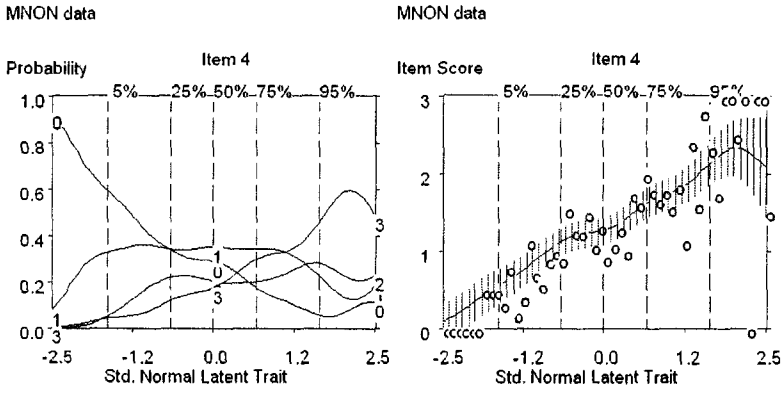


PAI Item #81

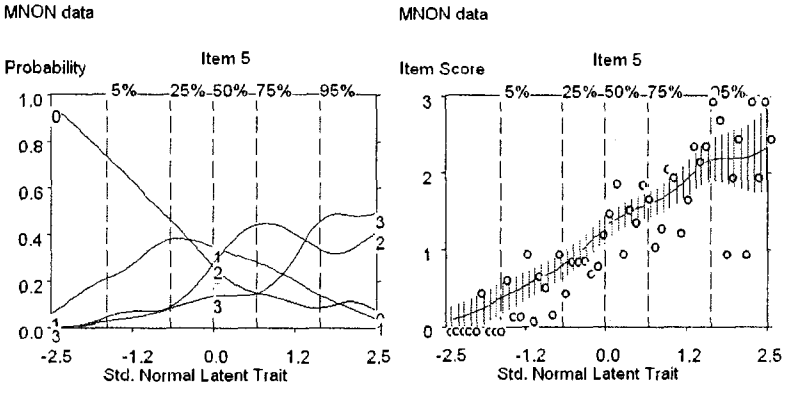




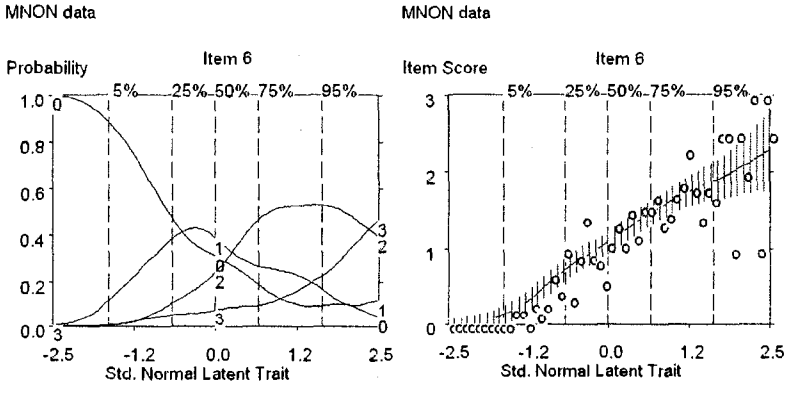
PAI Item #121



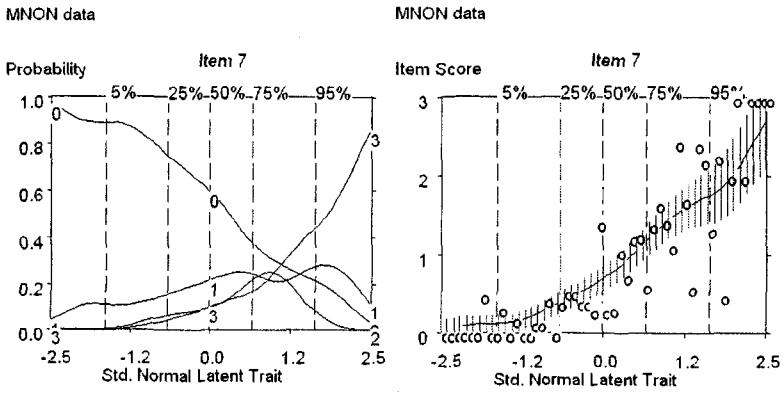
PAI Item #161



PAI Item #201



PAI Item #241



PAI Item #281

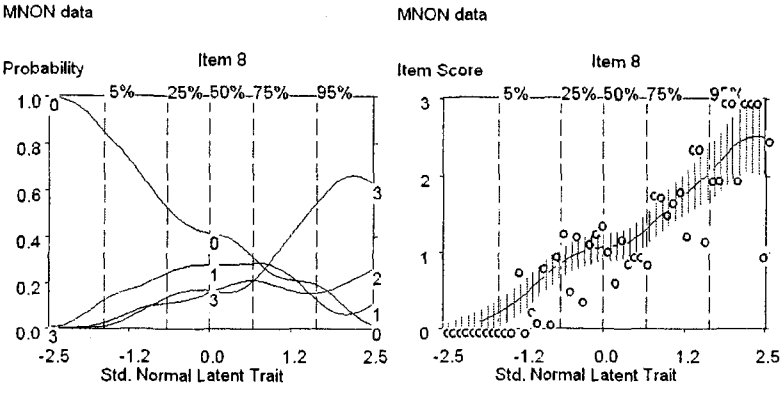


Figure 76. Option and item characteristic curves for the PAI scale NON based upon the responses of a sample of MMT patients (N = 323).

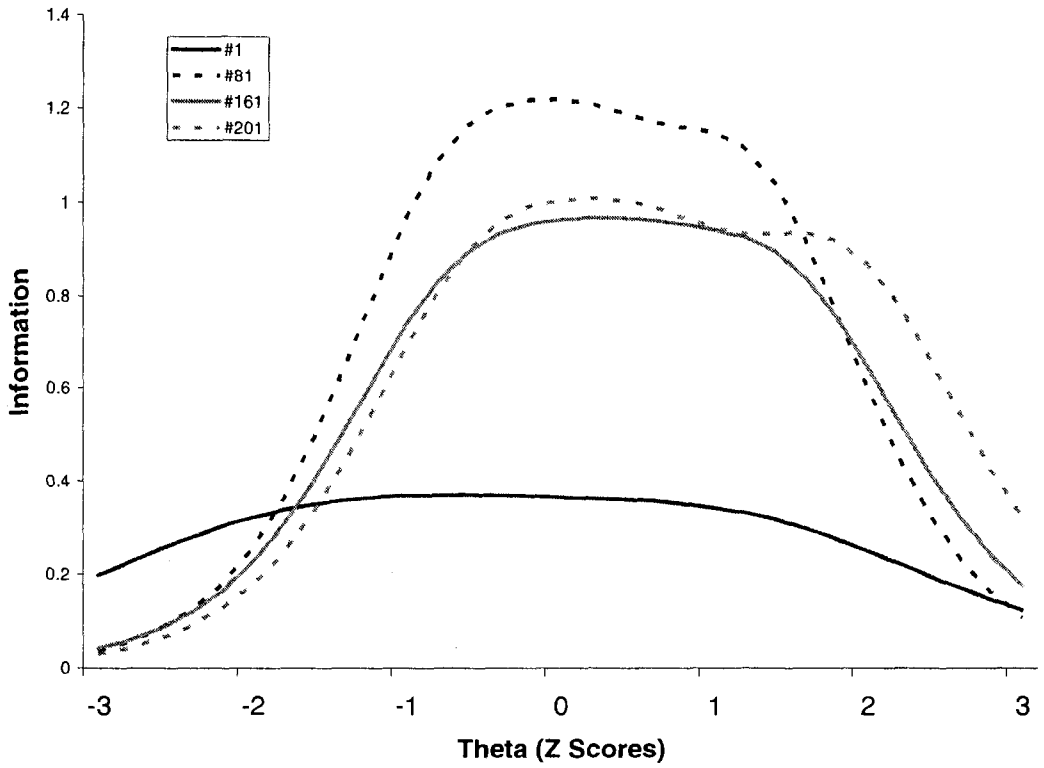
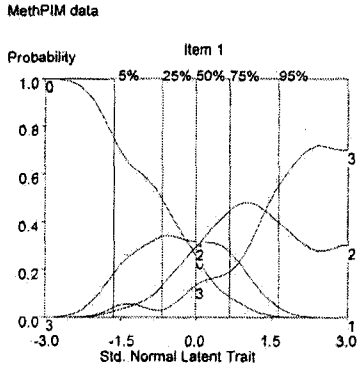
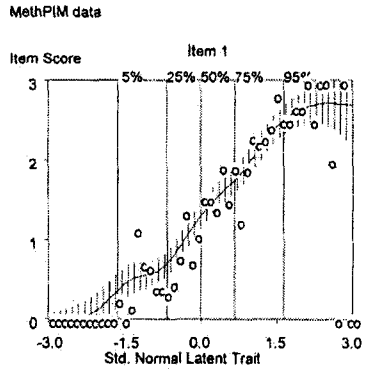


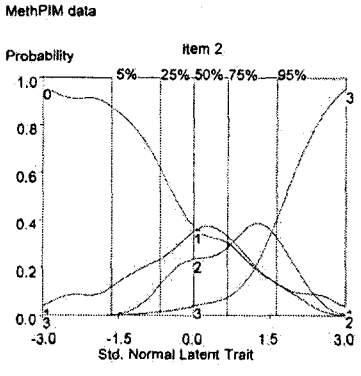
Figure 77. Item information curves for the four items of the NON-R scale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

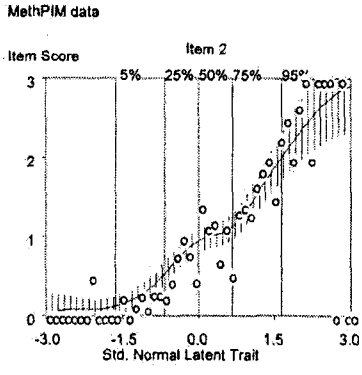
PAI Item #24



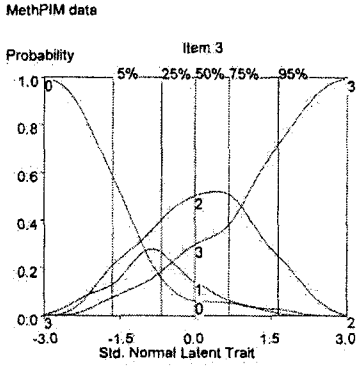


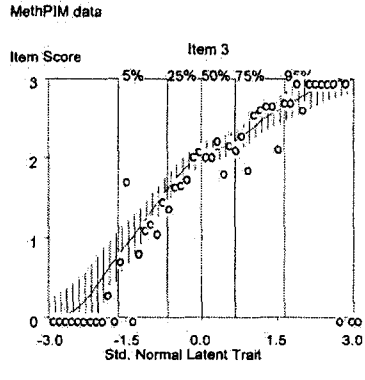
PAI Item #64



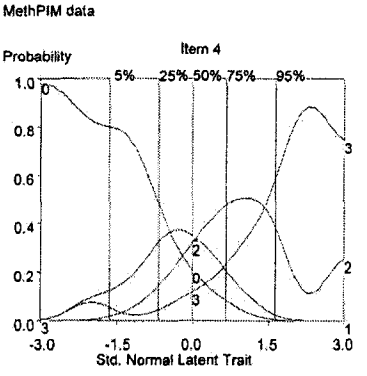


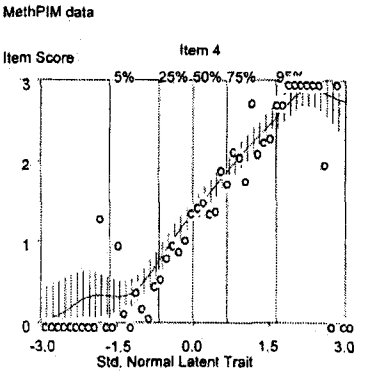
PAI Item #104



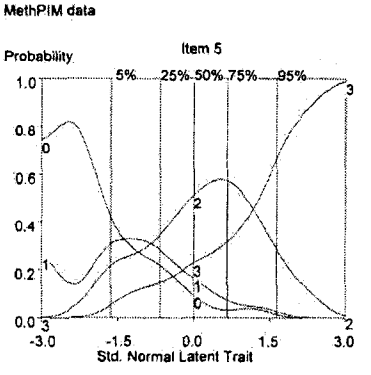


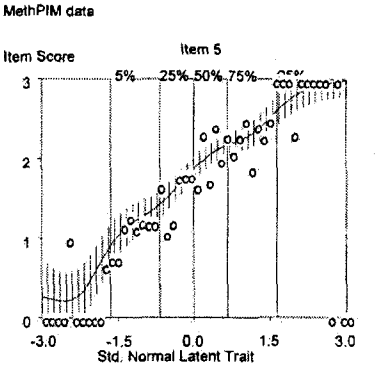
PAI Item #144



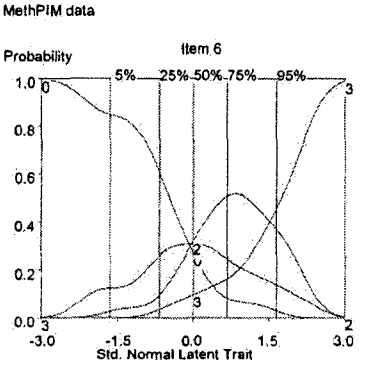


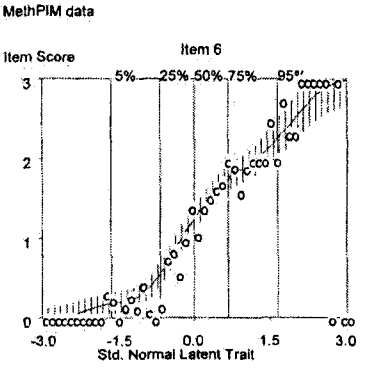
PAI Item #184



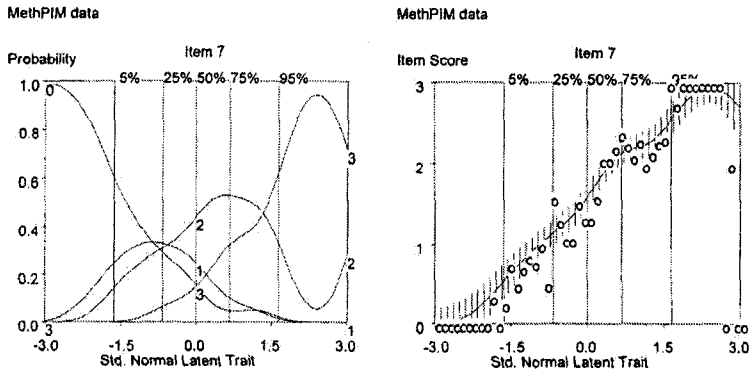


PAI Item #224

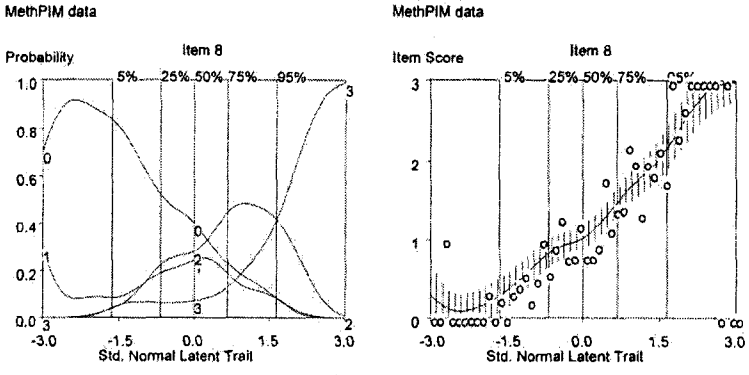




PAI Item #264



PAI Item #304



PAI Item #344

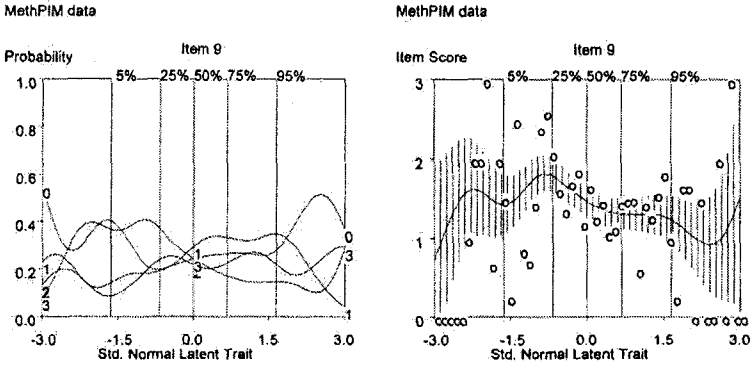


Figure 78. Option and item characteristic curves for the original PIM scale based upon responses to the PAI by a sample of MMT patients (N = 323).

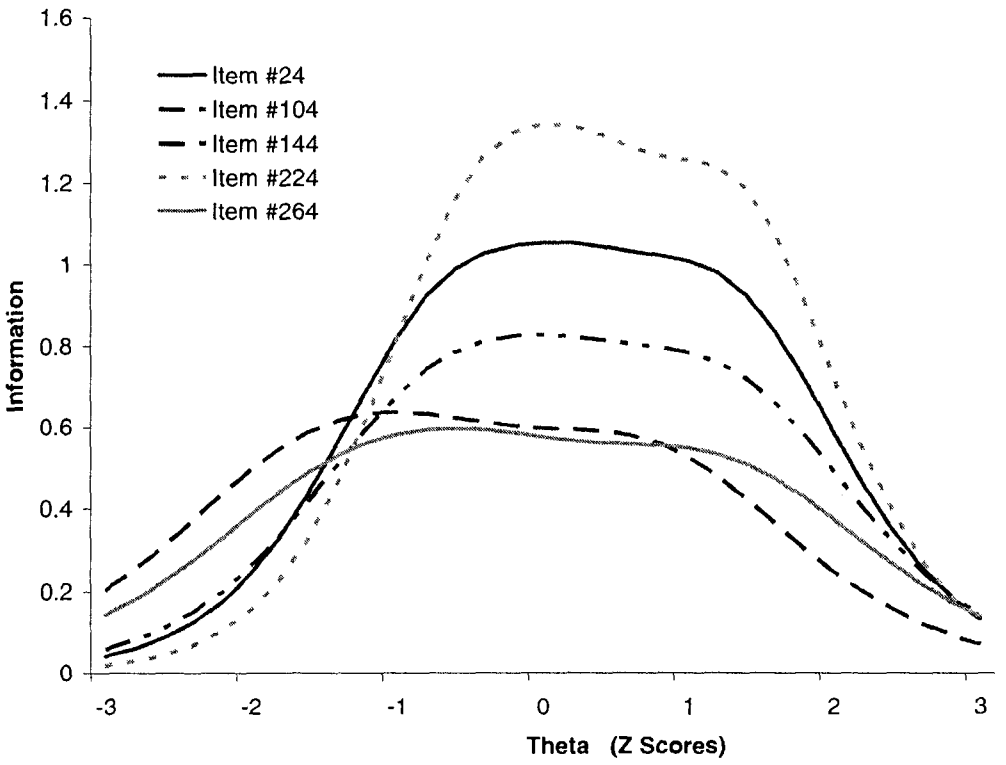


Figure 79. Item information curves for the final set of items for the PIM scale in use with a MMT population (N = 323).

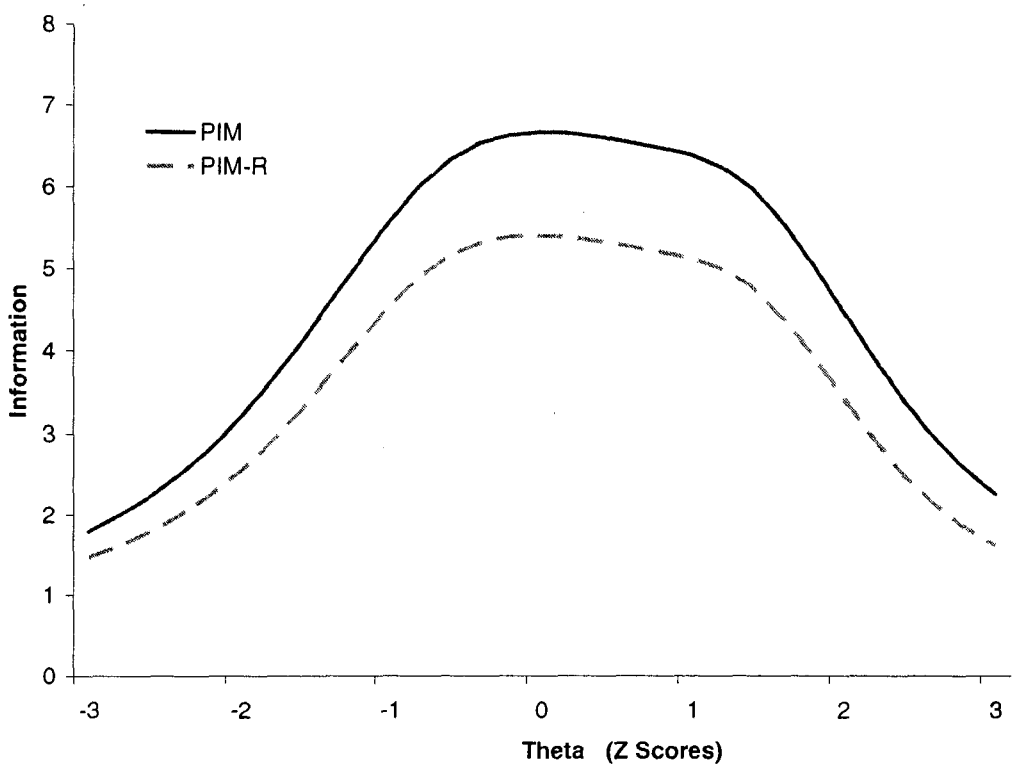
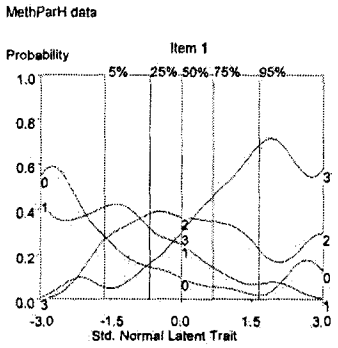
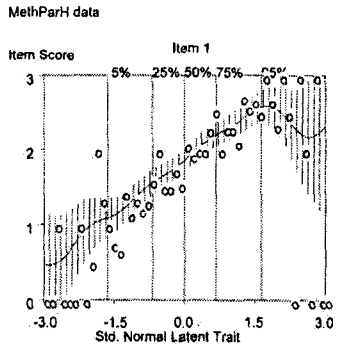


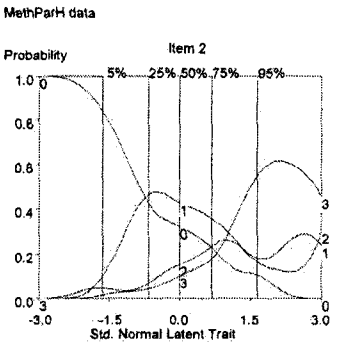
Figure 80. Total information curves for the PIM-R (5 items) versus initial estimates for the original PIM (9 items) (N = 323).

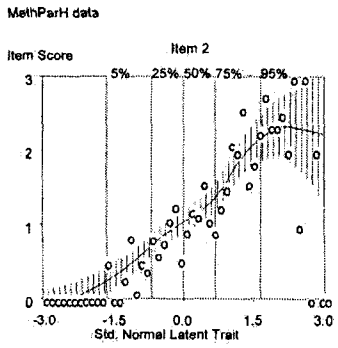
A. PAI Item #8



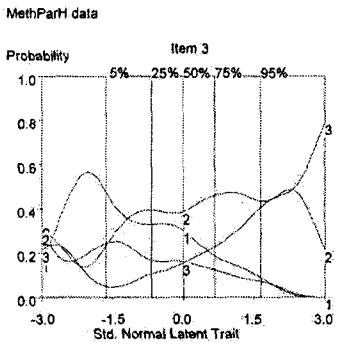


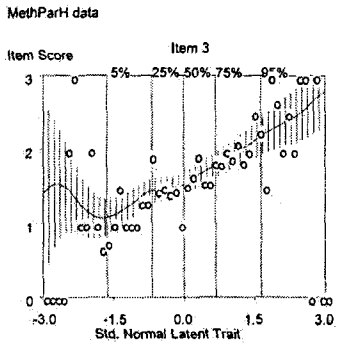
B. PAI Item #48



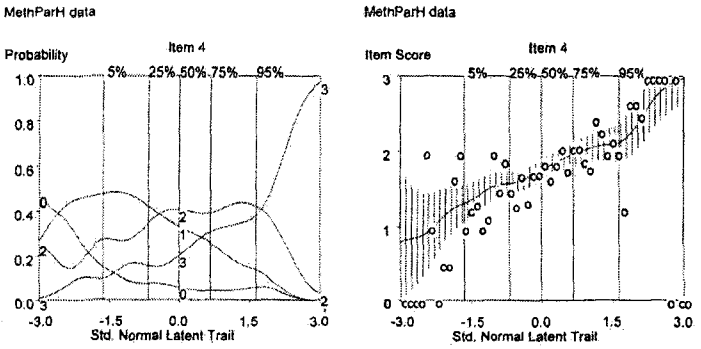


C. PAI Item #88

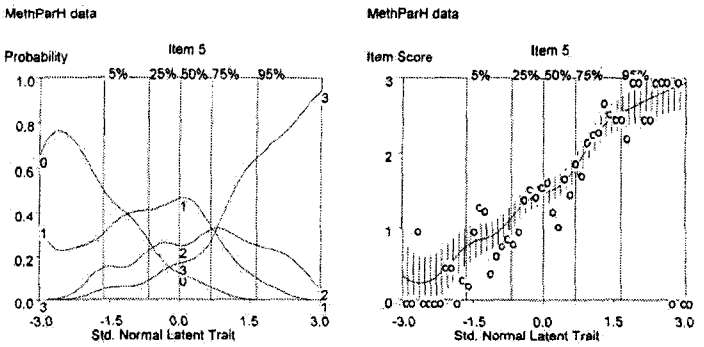




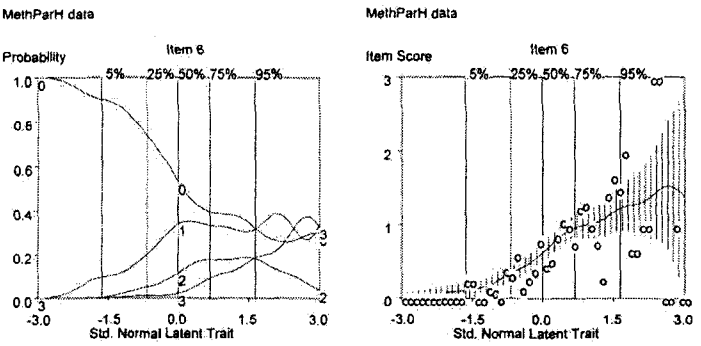
D. PAI Item #128



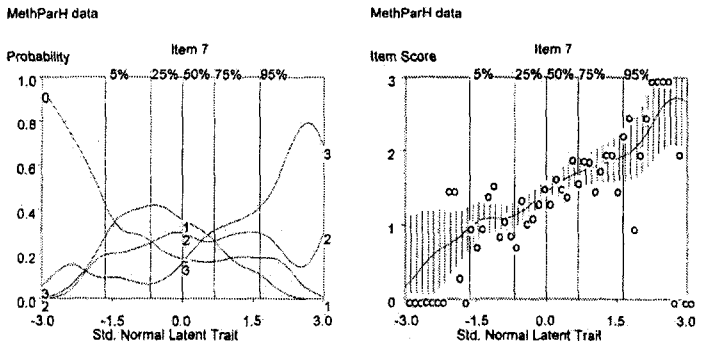
E. PAI Item #168



F. PAI Item #208



G. PAI Item #248



H. PAI Item #288

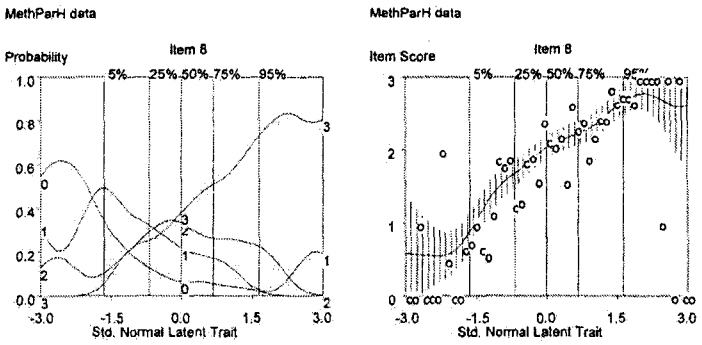


Figure 81. Option characteristic curves and Item characteristic Curves for the each item of the PAR-H subscale of the PAI for the methadone maintenance treatment population (N = 323).

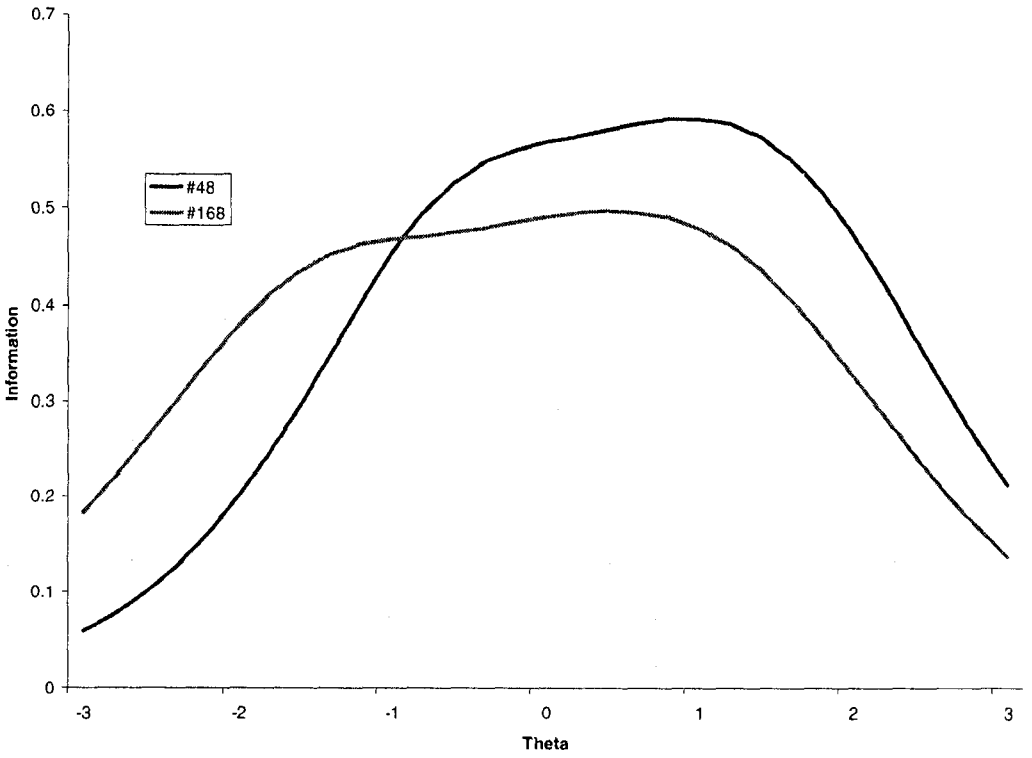
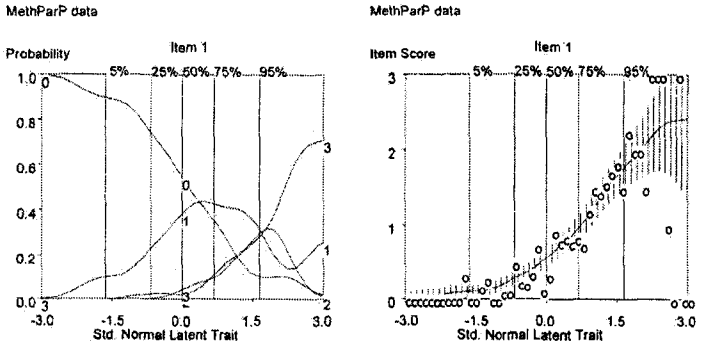
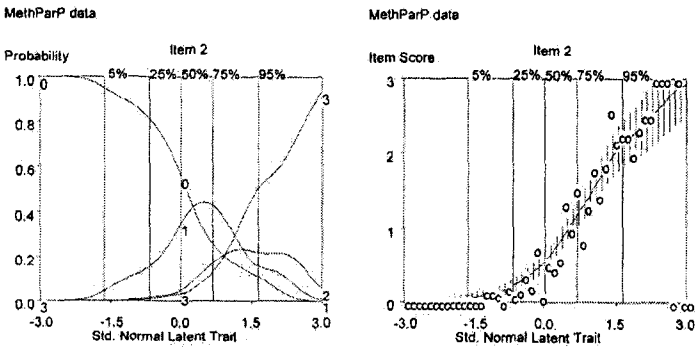


Figure 82. Information curves are displayed for each item of the PAR-H-R scale of the PAI, when administered to a sample of individuals receiving MMT ($N = 332$).

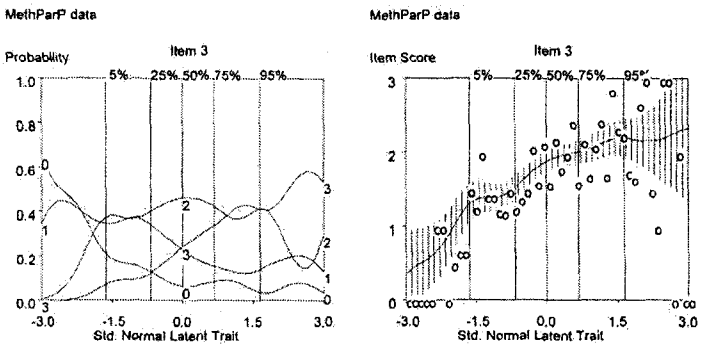
A. PAI Item #29



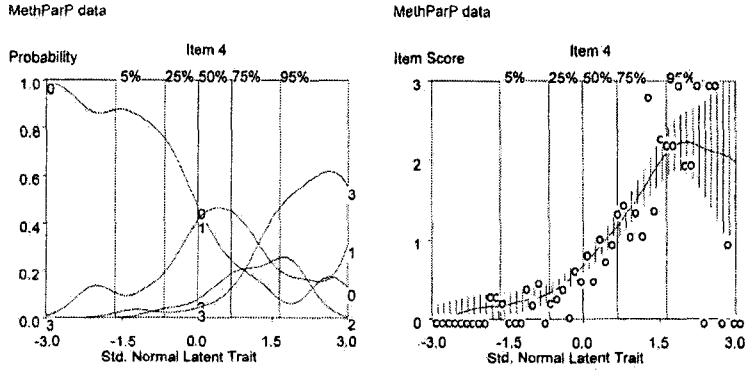
B. PAI Item #69



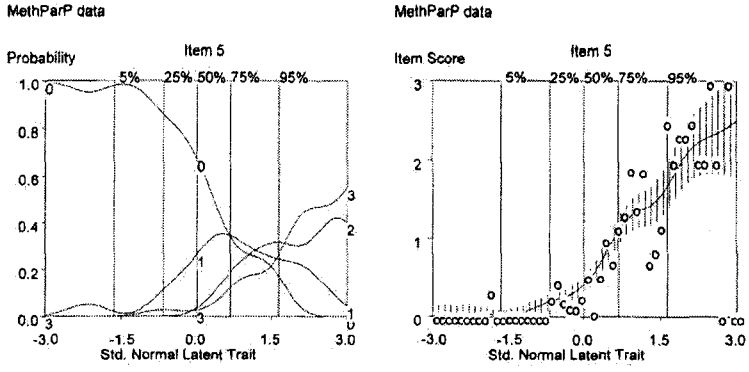
C. PAI Item#109



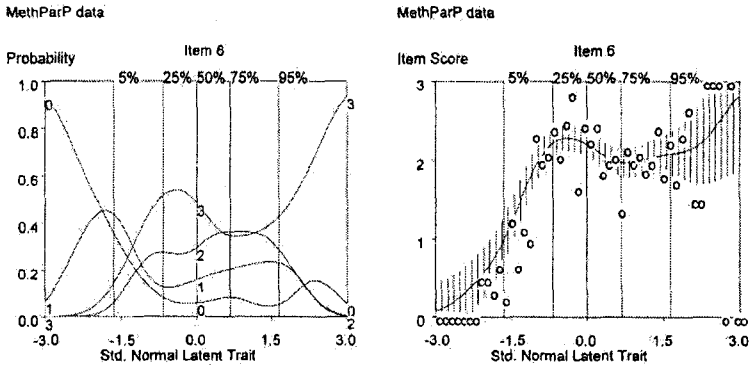
D. PAI Item #149



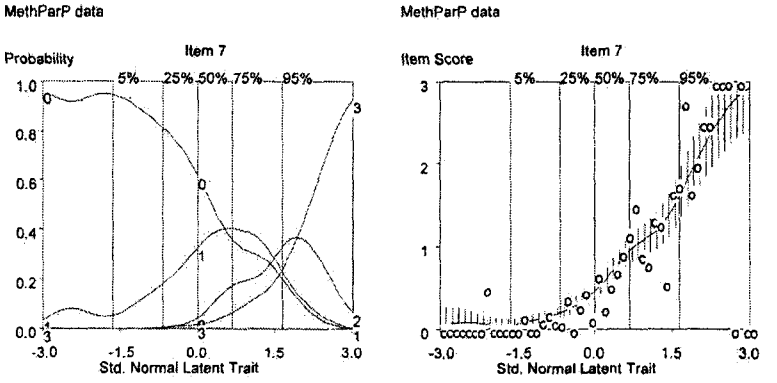
E. PAI Item #189



F. PAI Item #229



G. PAI Item #269



H. PAI Item #309

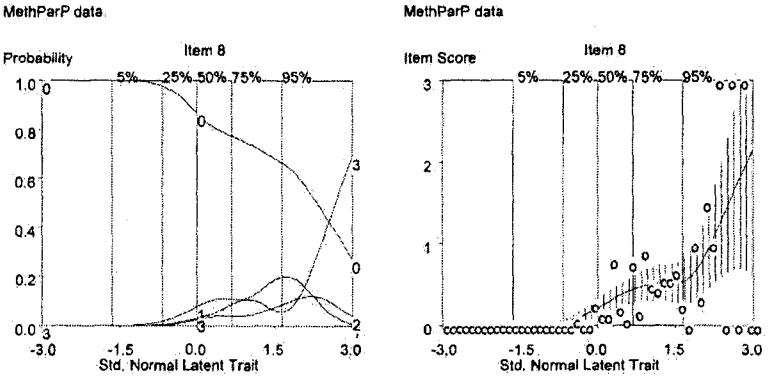


Figure 83. Option and item characteristic curves of the PAR-P subscale as plotted by the NIRT model software Testgraf which employs a Gaussian kernel smoothing technique. The plots are based upon item responses from a sample of MMT patients ($N = 323$).

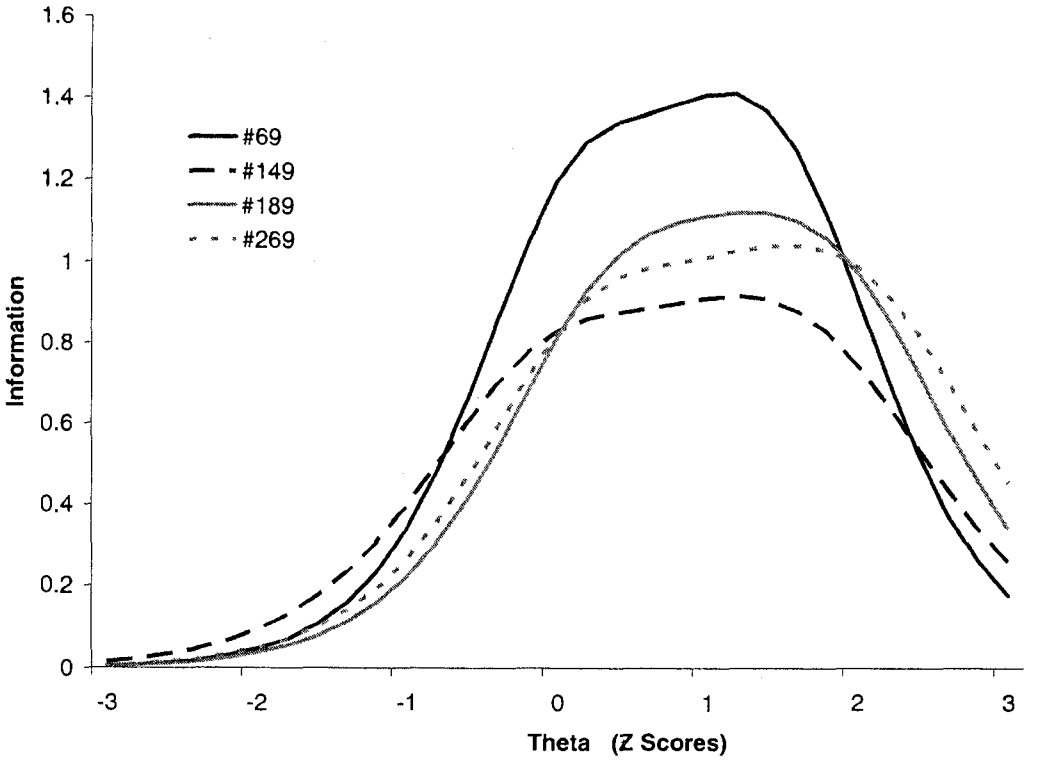


Figure 84. Item information curves for the PAR-P-R subscale based upon responses to the PAI from a sample of MMT patients ($N = 323$).

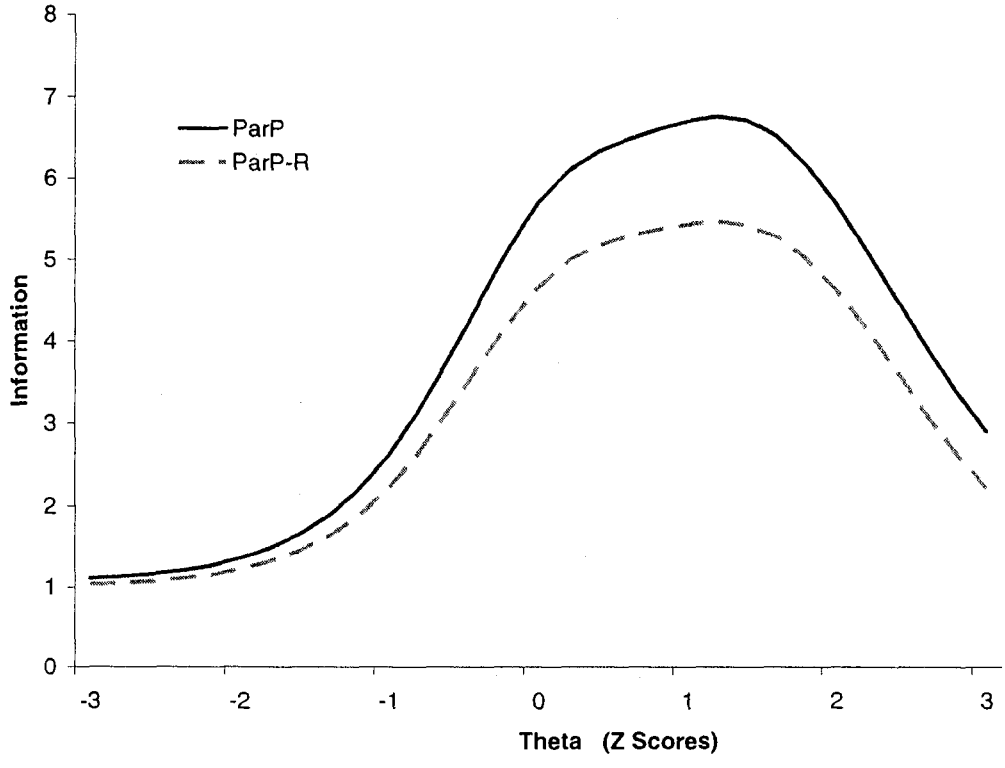
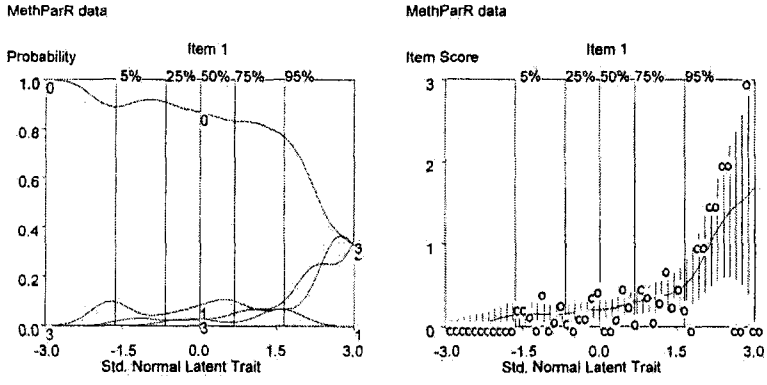
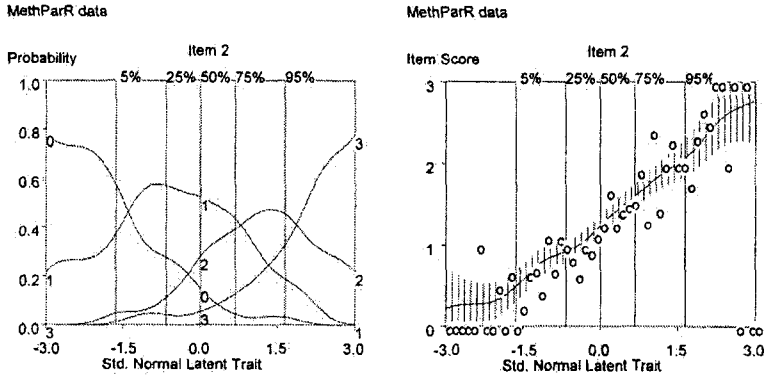


Figure 85. Total subscale information curves comparing the original eight item PAR-P subscale and the new four item PAR-P-R subscale based upon item responses from a sample of MMT patients ($N = 323$).

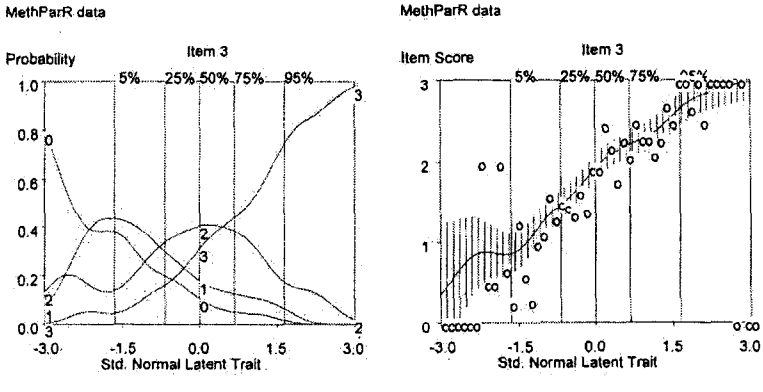
A. PAI Item #37



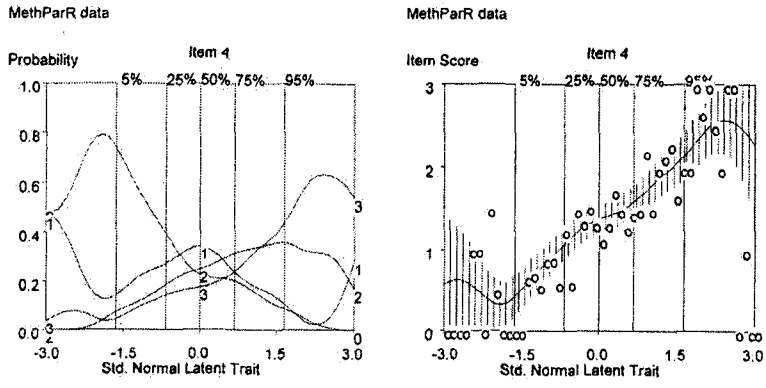
B. PAI Item #77



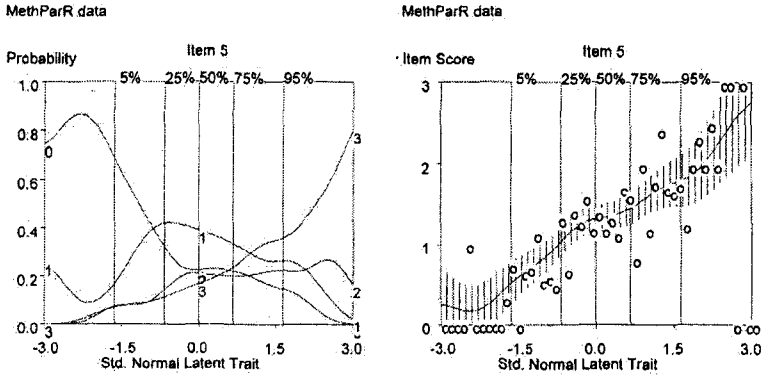
C. PAI Item #117



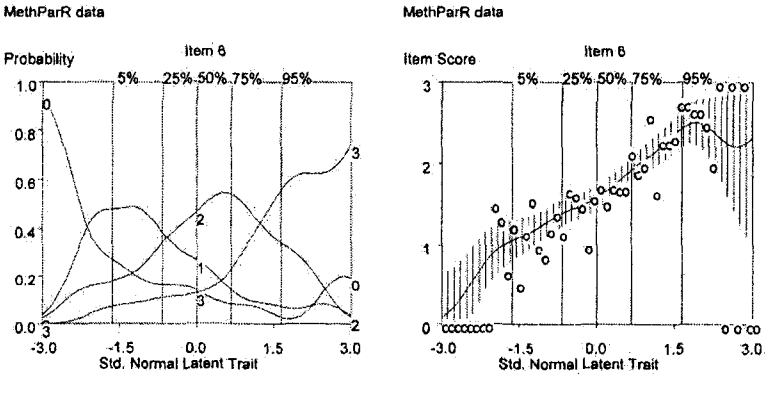
D. PAI Item #157



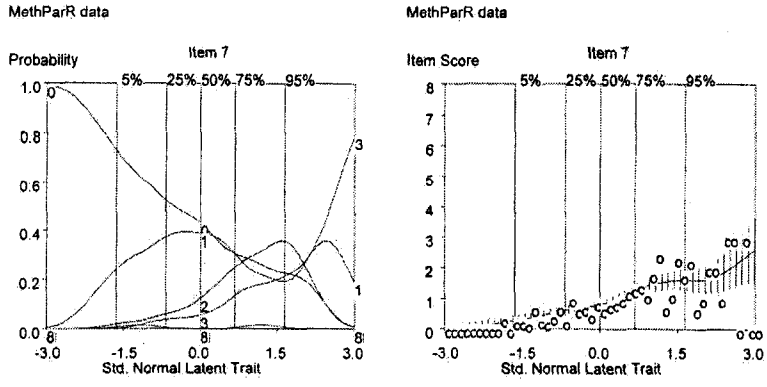
E. PAI Item #197



F. PAI Item #237



G. PAI Item #277



H. PAI Item #317

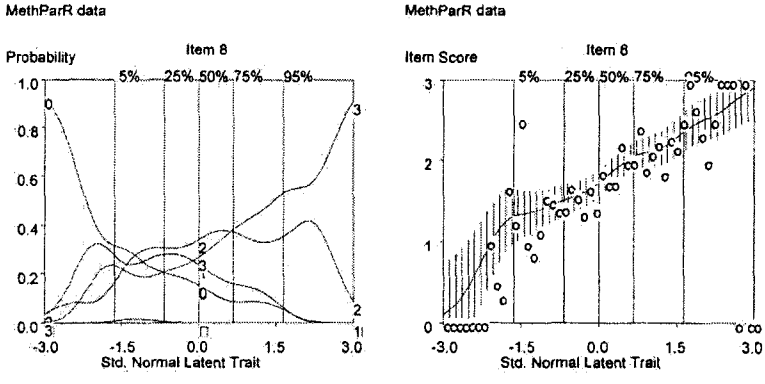


Figure 86. Option and item characteristic curves for the PAR-R subscale of the PAI based upon responses from a MMT sample ($N = 323$).

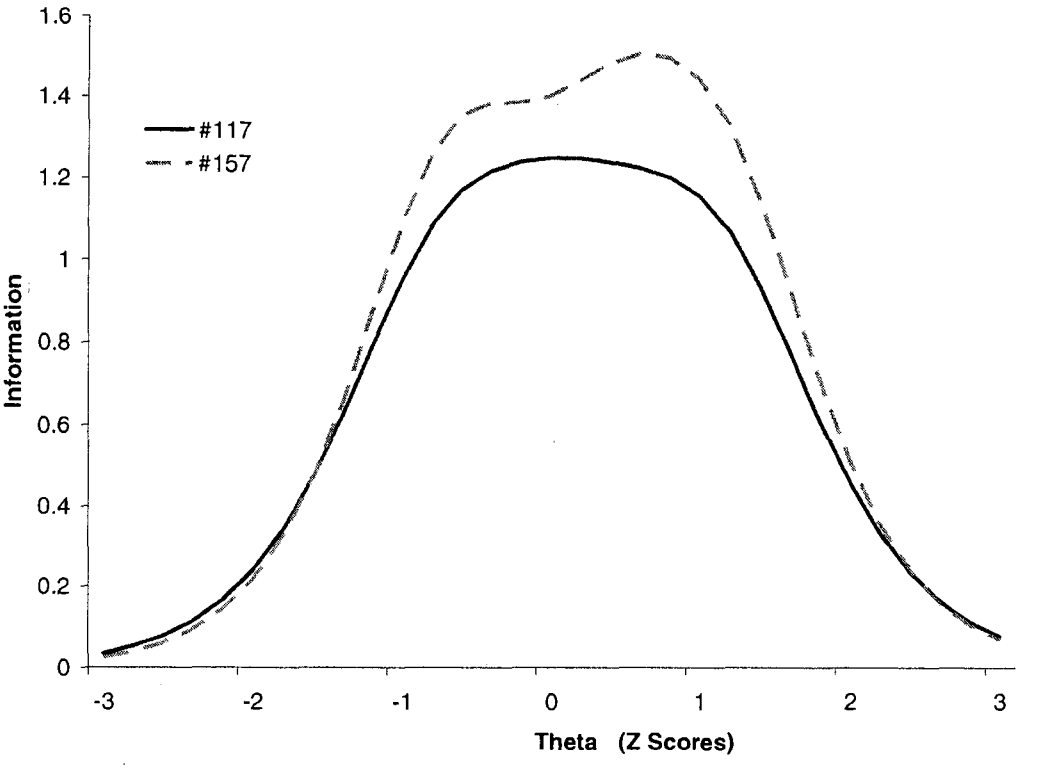
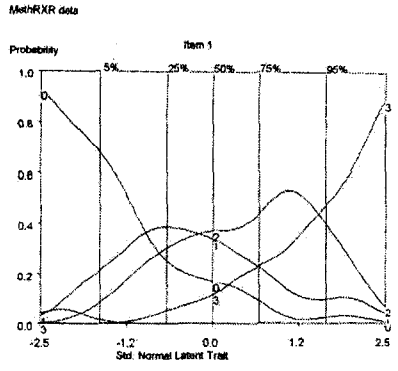
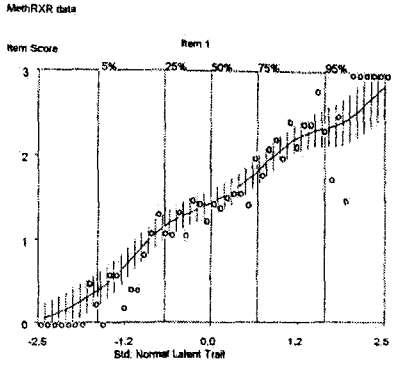


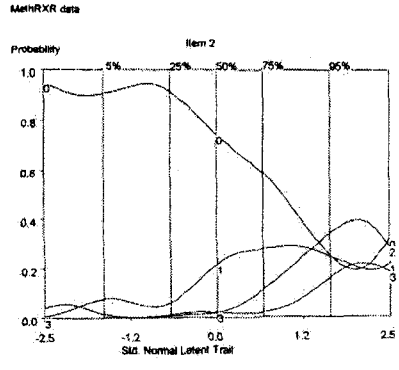
Figure 87. Item information curves for the two items of the PAR-R-R subscale based upon responses to the PAI by a sample of MMT patients (N = 323).

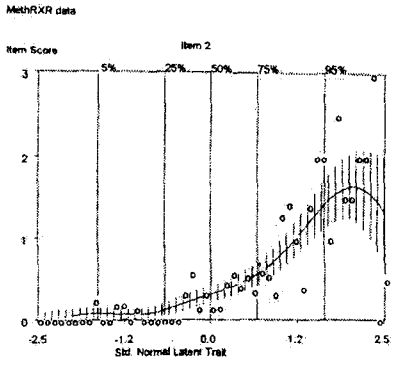
PAI Item #2



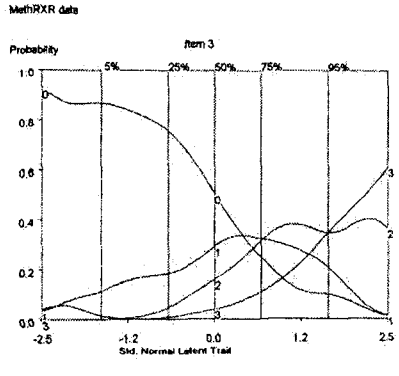


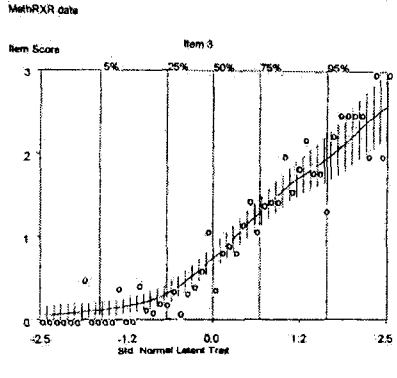
PAI Item #42



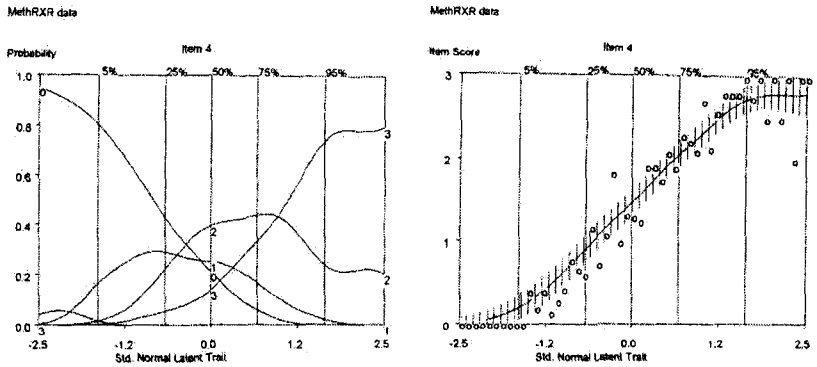


PAI Item #82

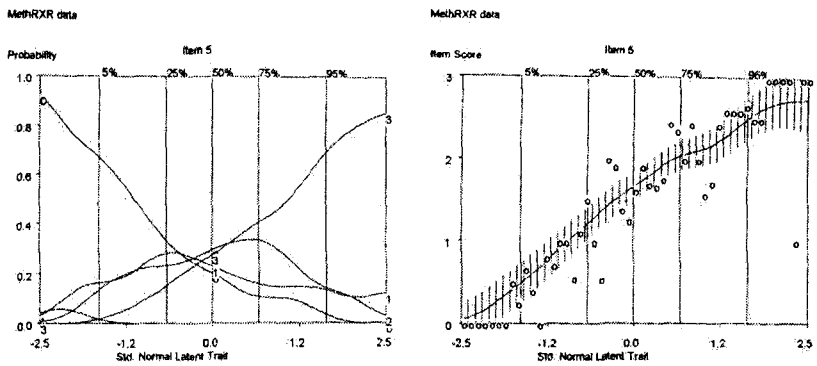




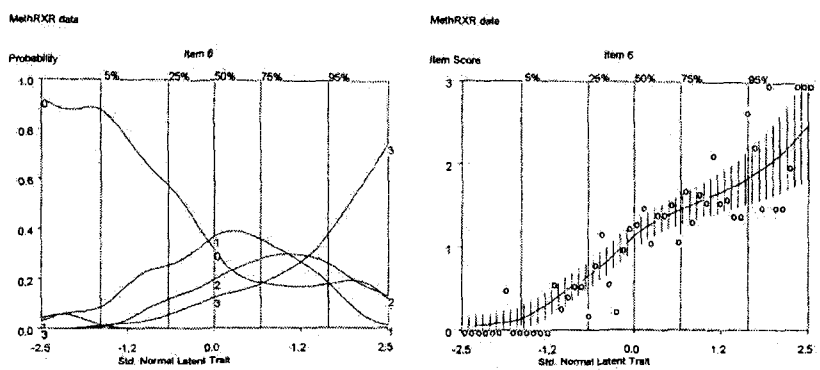
PAI Item #122



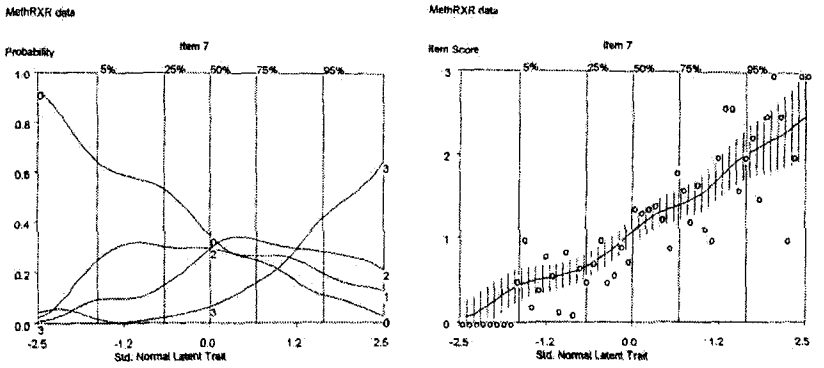
PAI Item #162



PAI Item #202



PAI Item #242



PAI Item #282

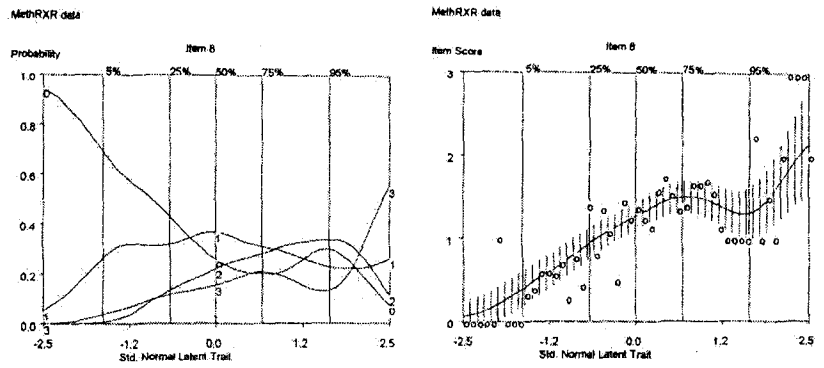


Figure 88. Option and item characteristic curves for the items from the RXR scale of the PAI based upon responses from a MMT sample (N = 323).

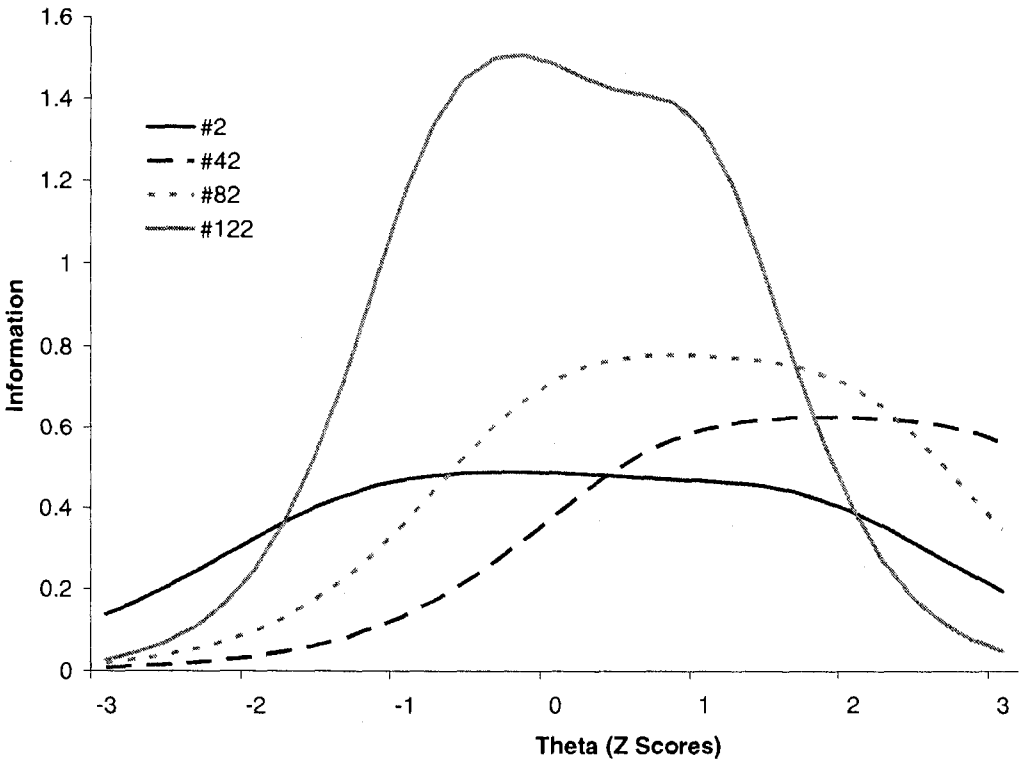


Figure 89. Item information curves for the four items of the RXR-R scale based upon responses of a sample of MMT patients ($N = 323$).

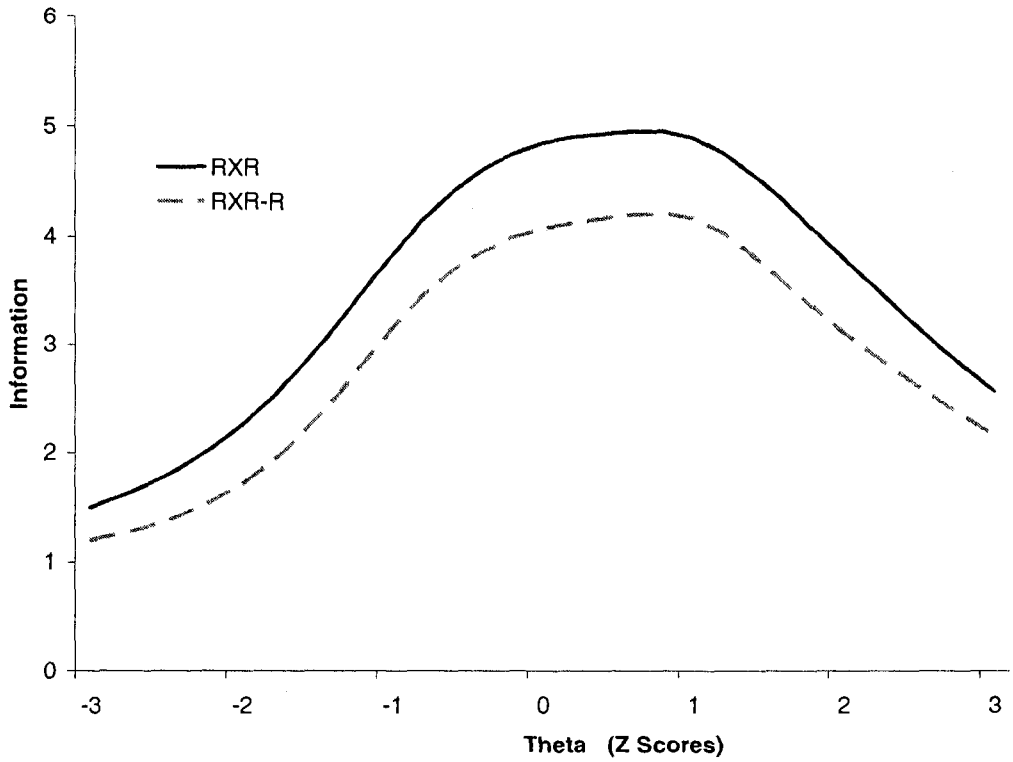
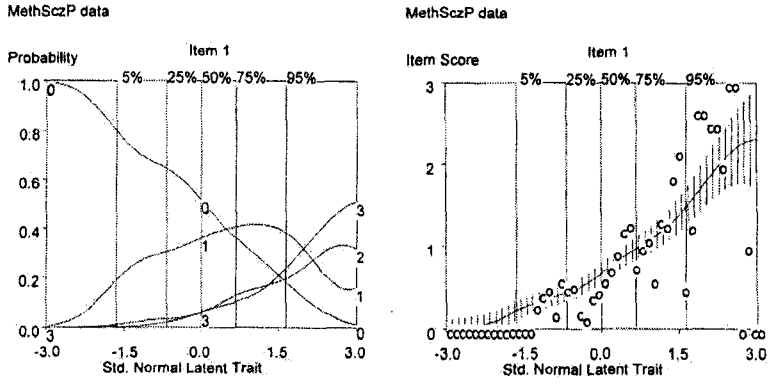
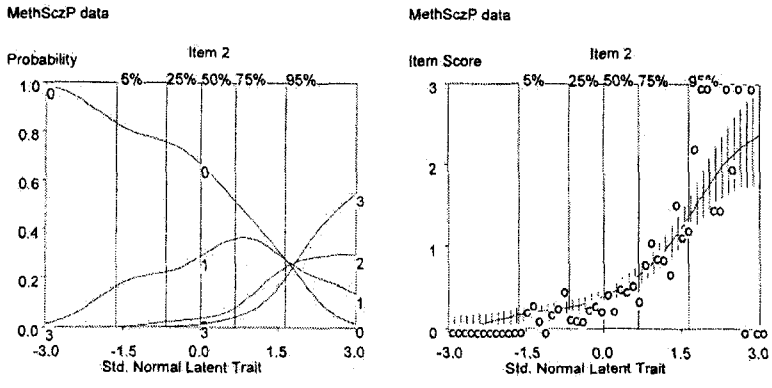


Figure 90. Total information curves comparing the original RXR scale with the RXR-R scale based upon responses from a sample of MMT patients ($N = 323$).

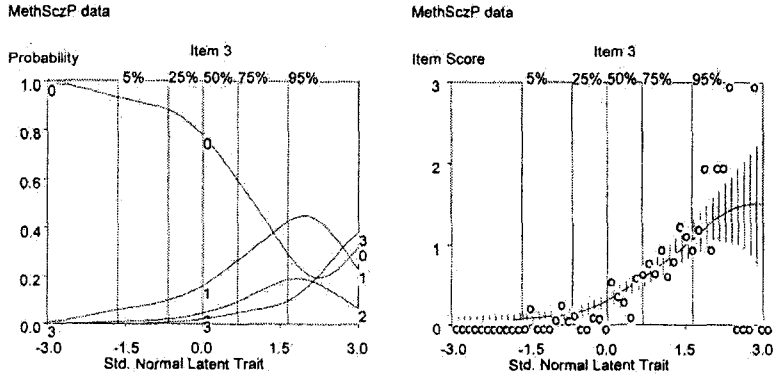
PAI Item #10



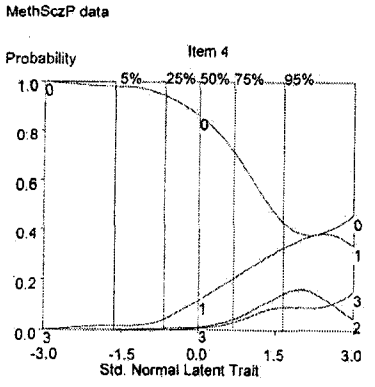
PAI Item #50

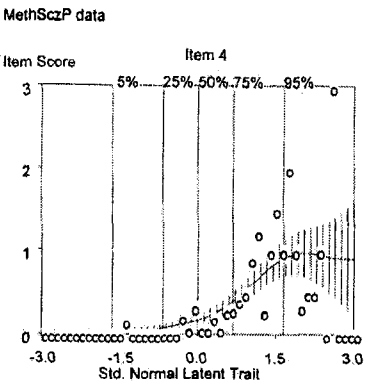


PAI Item #90

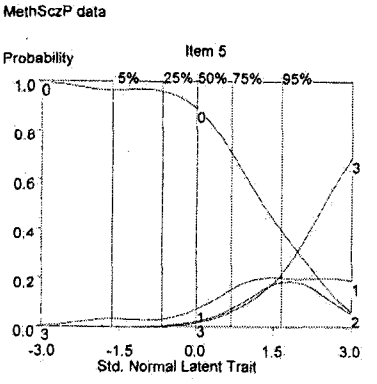


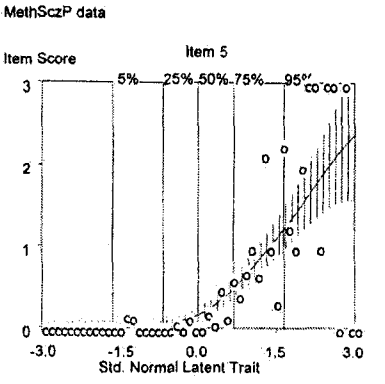
PAI Item #130



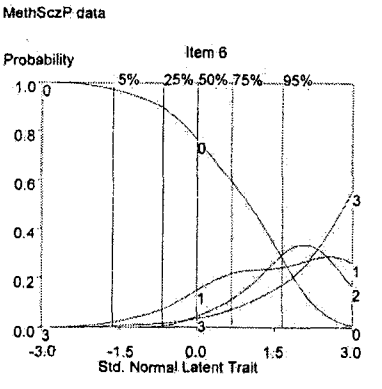


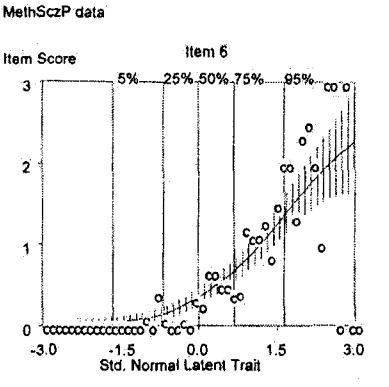
PAI Item #170



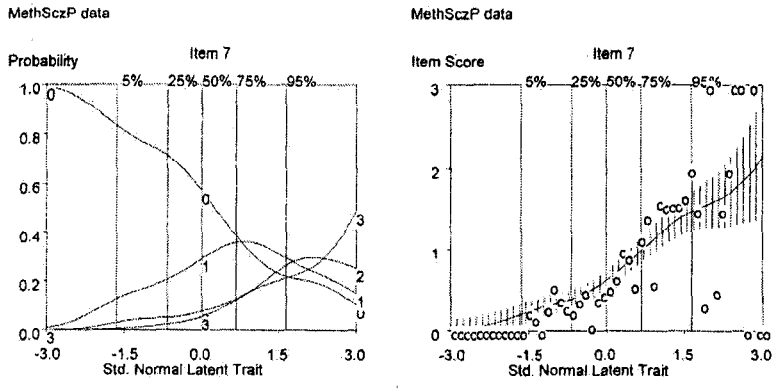


PAI Item #210





PAI Item #250



PAI Item #290

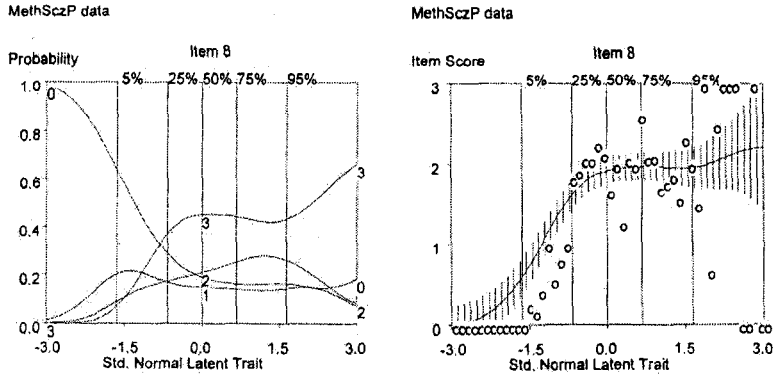


Figure 91. Option characteristic curves (OCC) and Item characteristic Curves (ICC) for the each item of the SCZ-P subscale of the PAI for the methadone maintenance treatment population (N = 323).

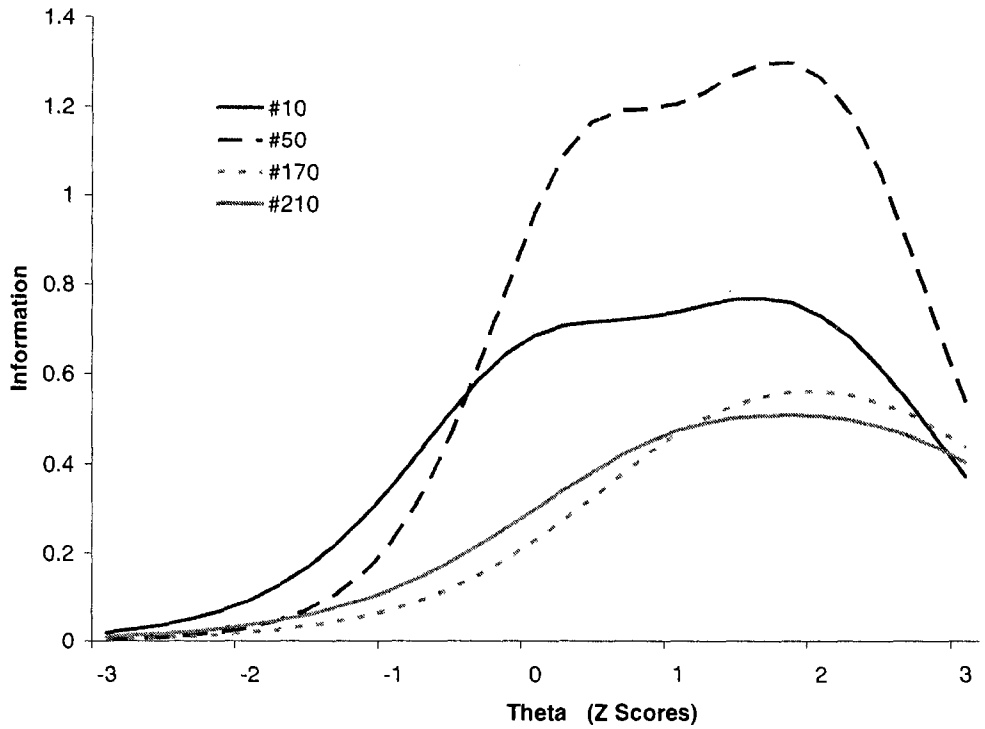


Figure 92. Item information curves for the final set of items for the SCZ-P-R scale in use with a MMT population ($N = 323$).

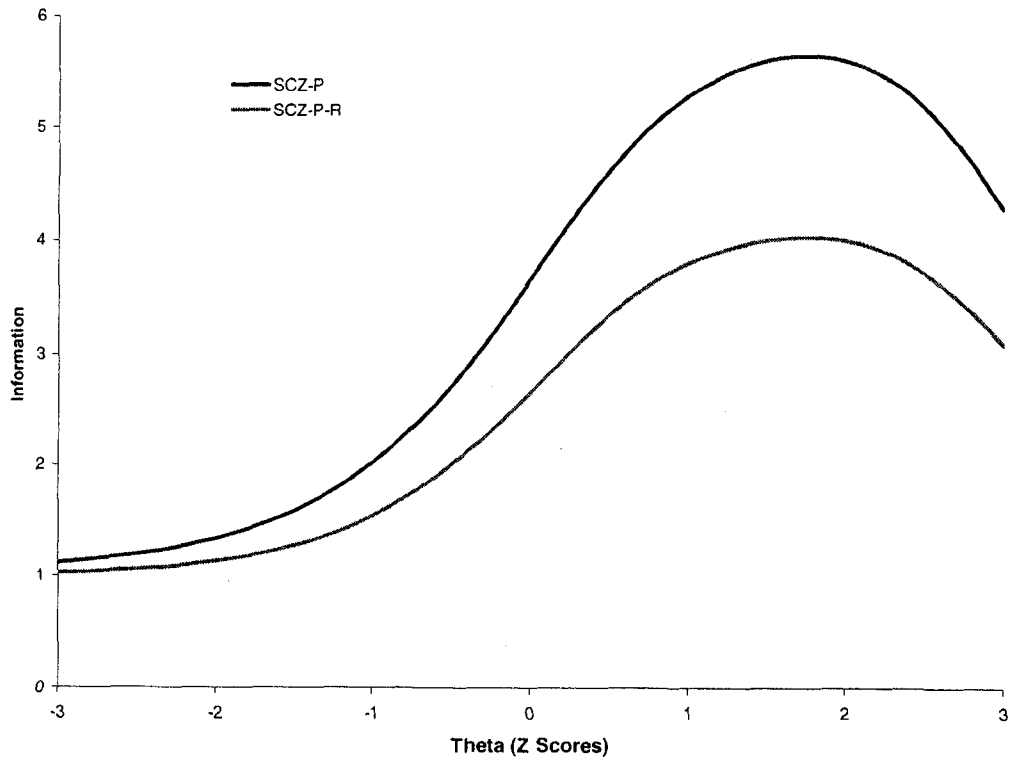
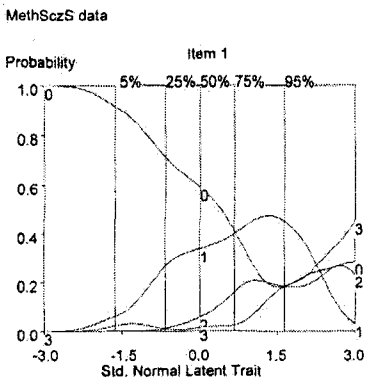
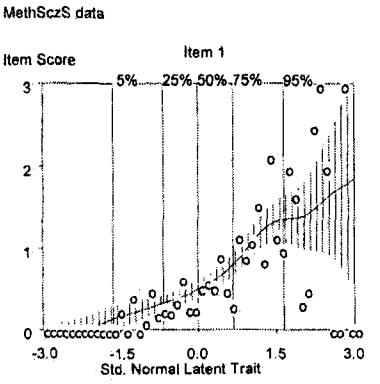


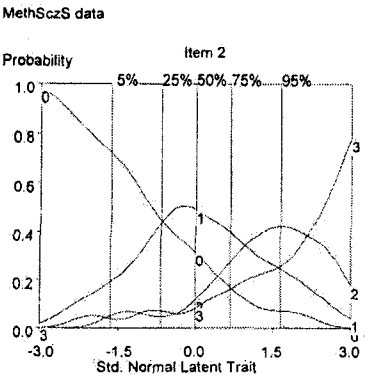
Figure 93. Total information curves for the SCZ-P-R (4 items) versus initial estimates for the original SCZ P (8 items) in a sample of MMT patients (N = 323).

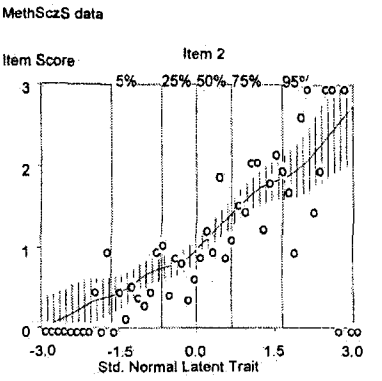
PAI Item #30



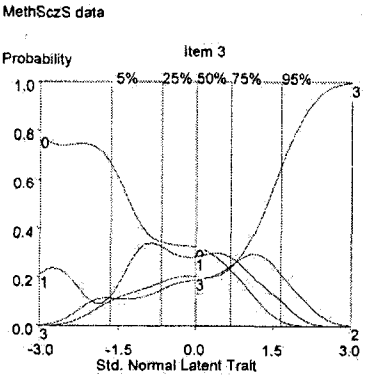


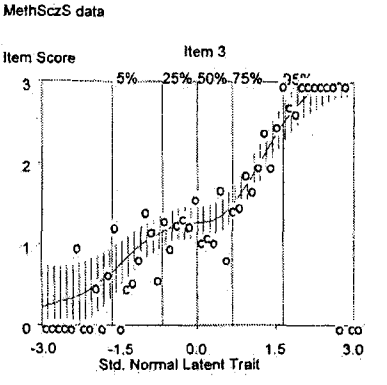
PAI Item #70



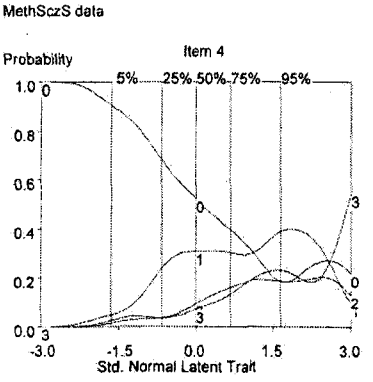


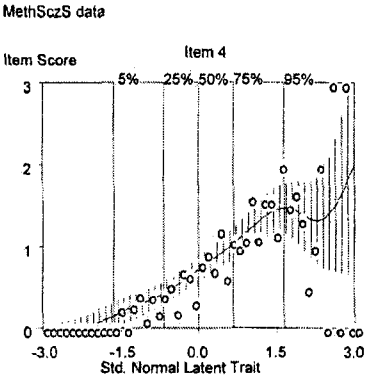
PAI Item #110



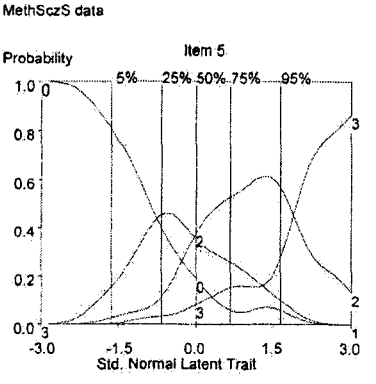


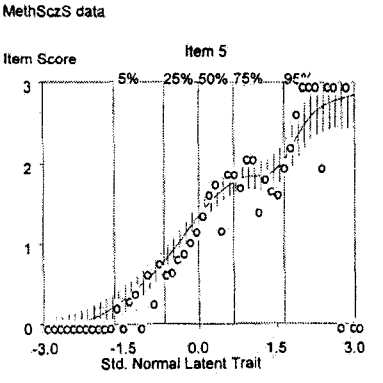
PAI Item #150



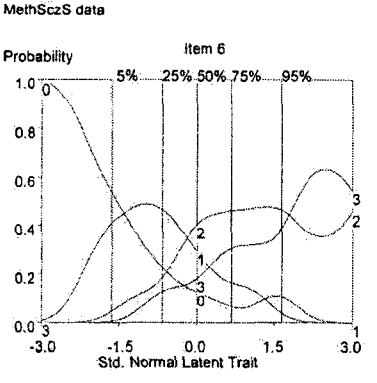


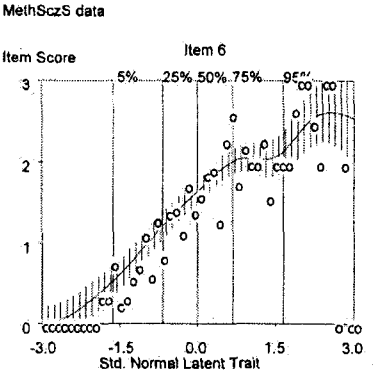
PAI Item #190



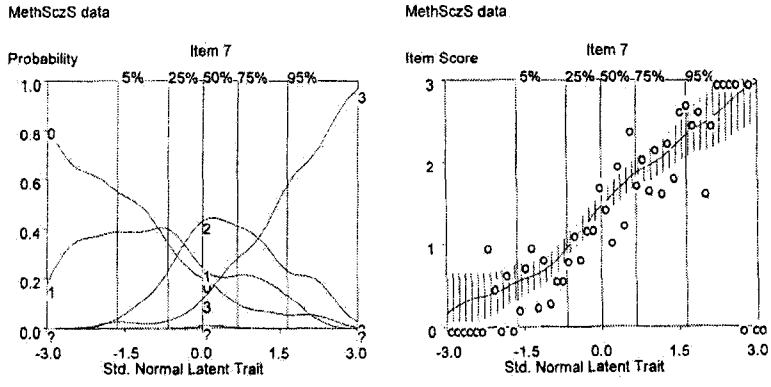


PAI Item #230





PAI Item #270



PAI Item #310

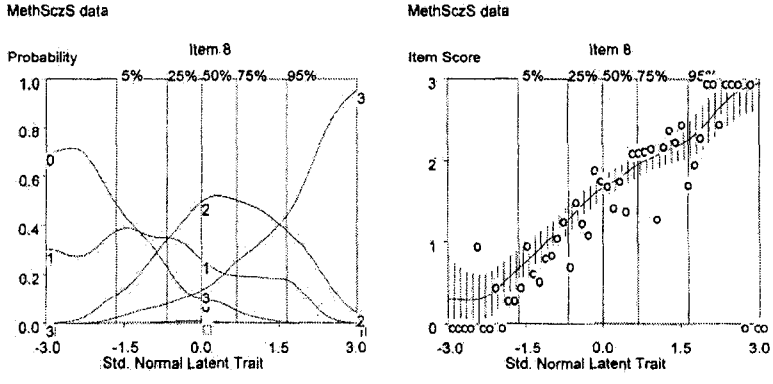


Figure 94. Option characteristic curves (OCC) and Item characteristic Curves (ICC) for the each item of the SCZ S scale of the PAI for the methadone maintenance treatment population (N = 323).

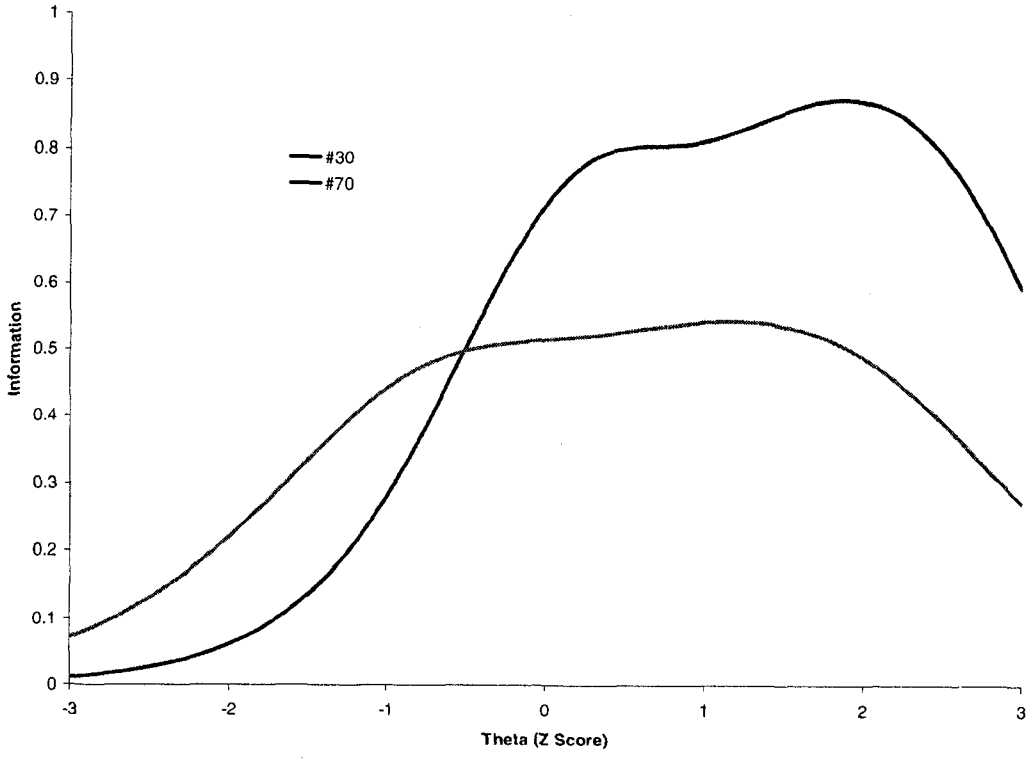
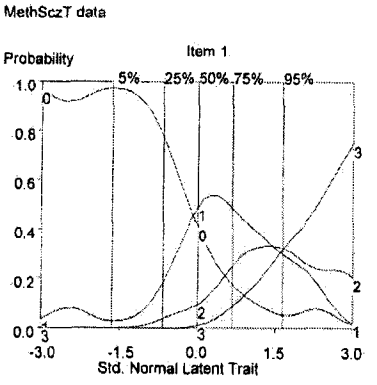
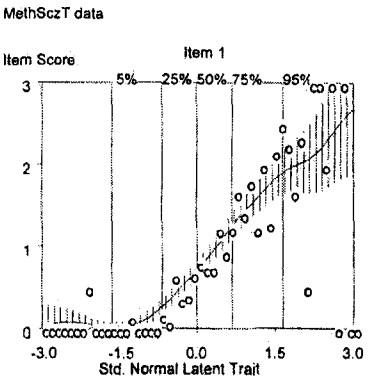


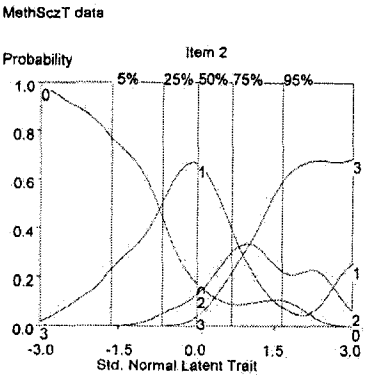
Figure 95. Item information curves for the final set of items for the SCZ-S-R scale in use with a MMT population (N = 323).

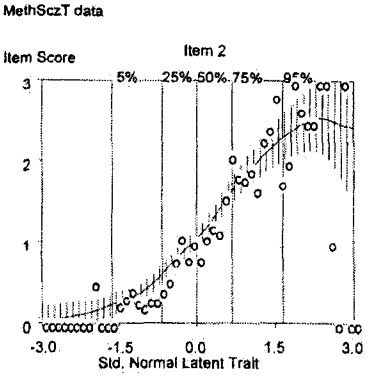
PAI Item #38



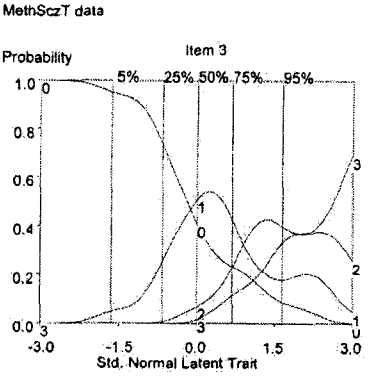


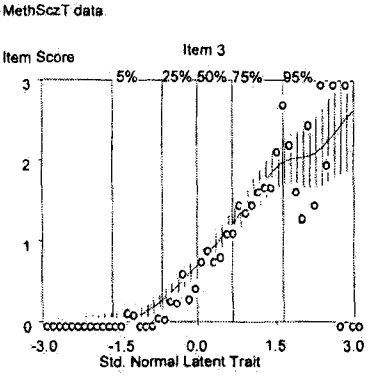
PAI Item #78



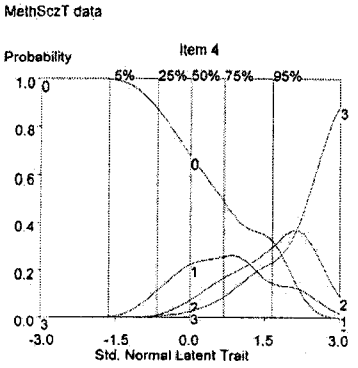


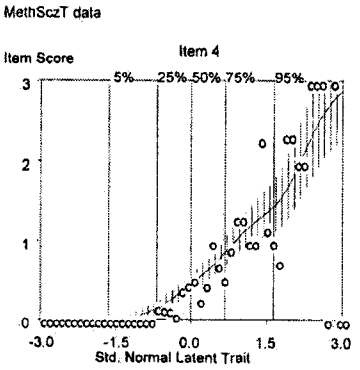
PAI Item #118



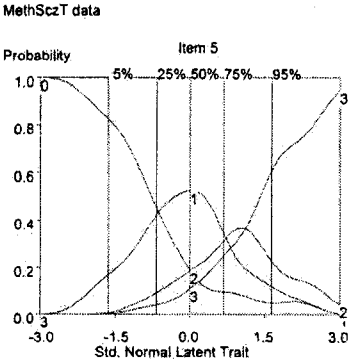


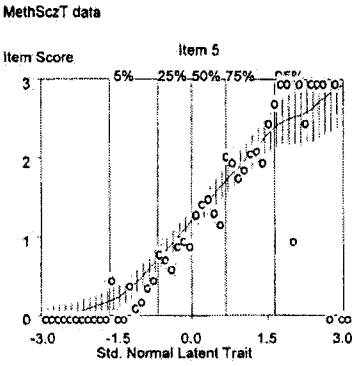
PAI Item #158



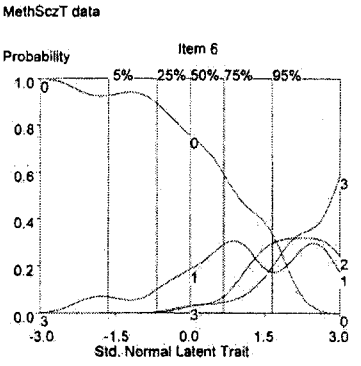


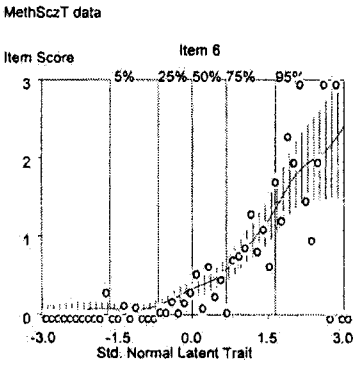
PAI Item #198



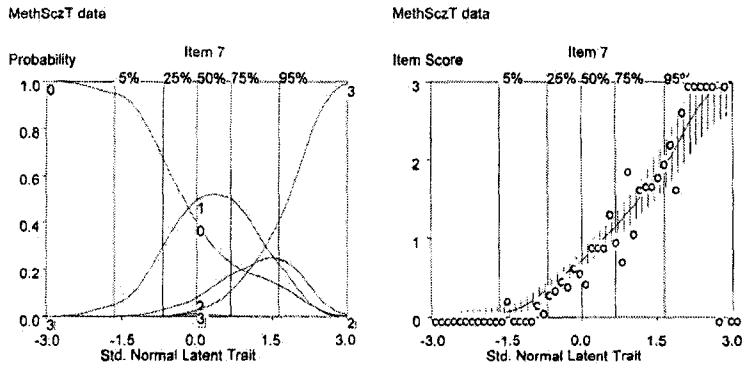


PAI Item #238





PAI Item #278



PAI Item #318

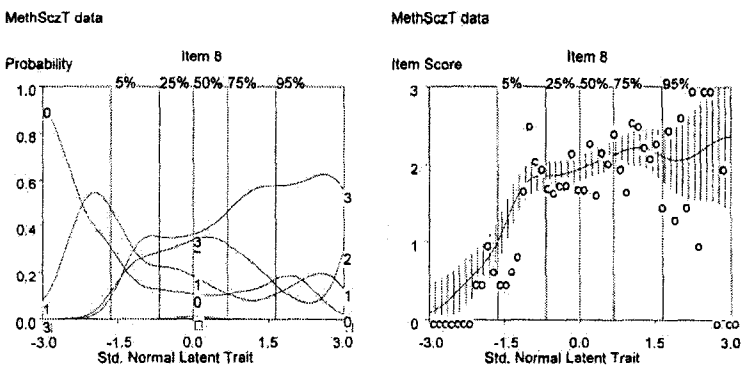


Figure 96. Option characteristic curves and item characteristic Curves for each item of the SCZ T scale of the PAI for the methadone maintenance treatment population (N = 323).

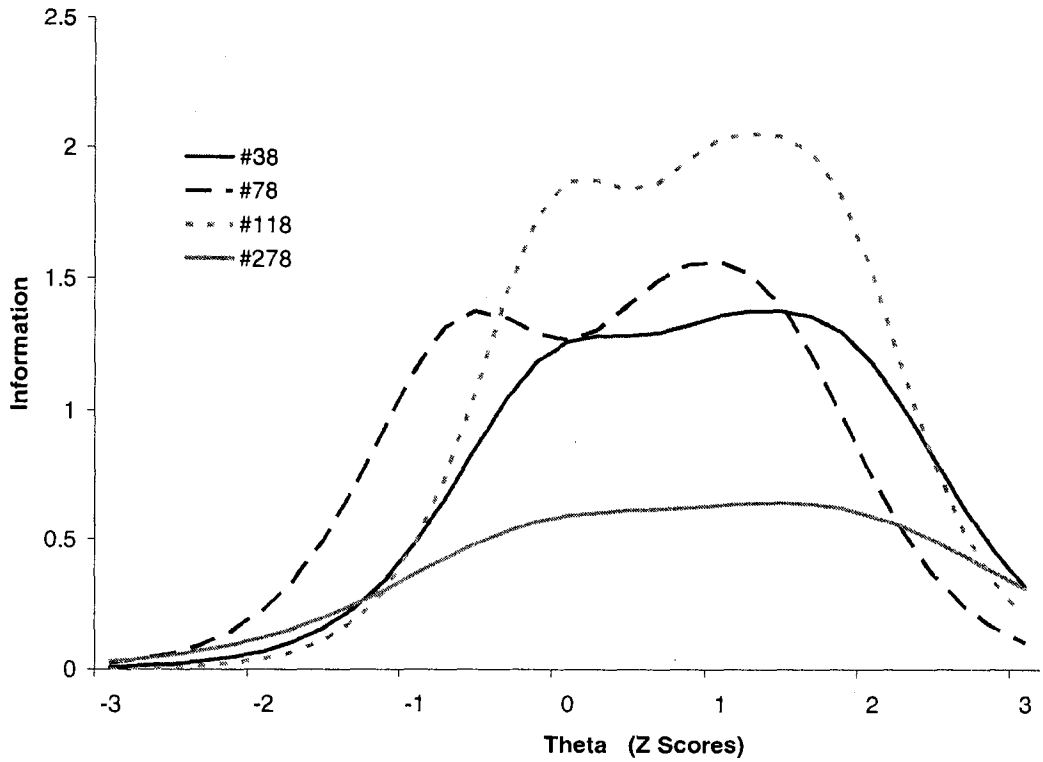


Figure 97. Item information curves for the final set of items for the SCZ-T-R scale in use with a MMT population ($N = 323$).

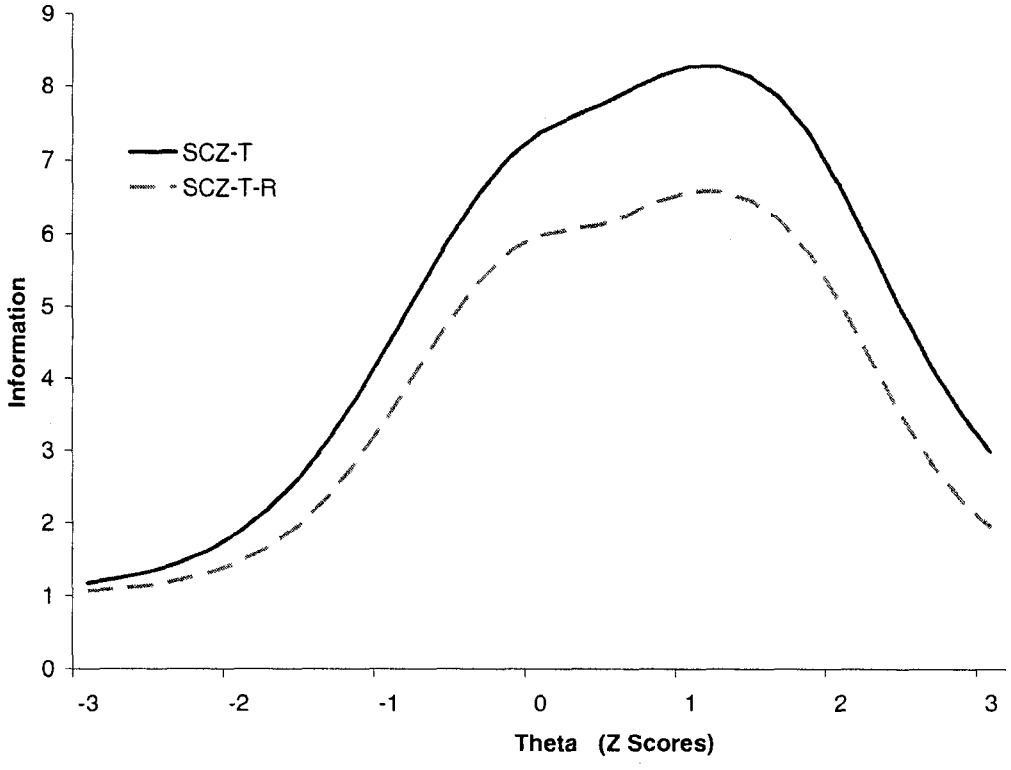
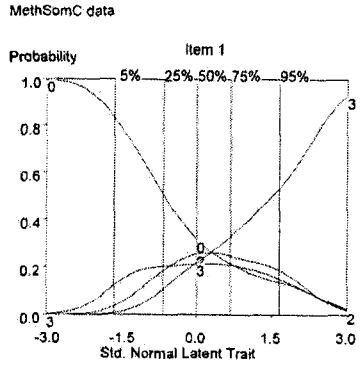
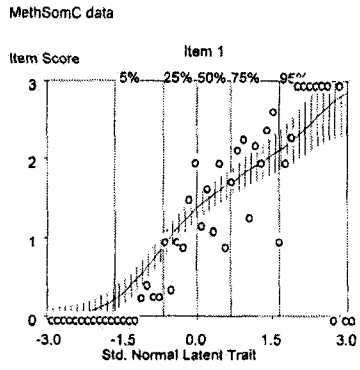


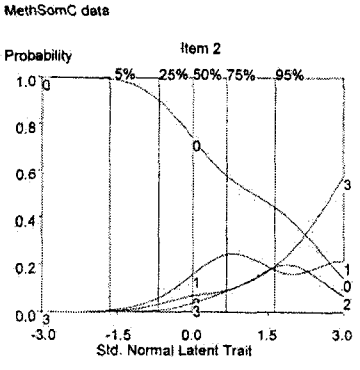
Figure 98. Total information curves for the SCZ-T-R (4 items) versus initial estimates for the original SCZ-T (8 items) in a sample of MMT patients (N = 323).

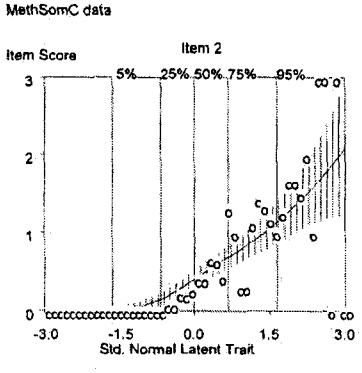
PAI Item #3



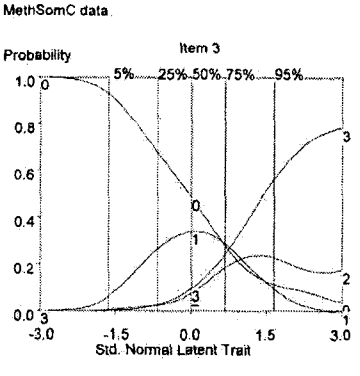


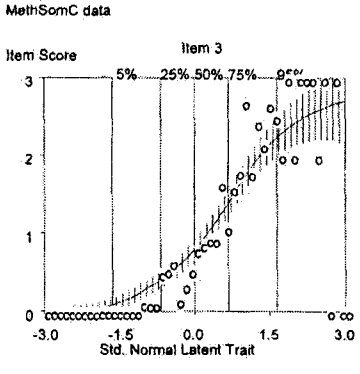
PAI Item #43



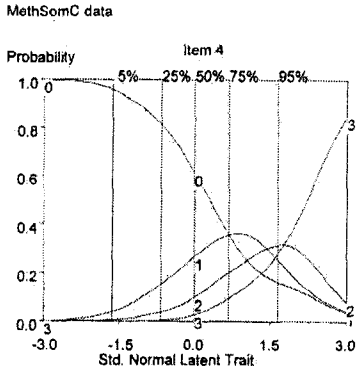


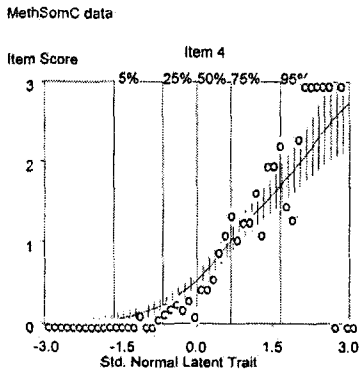
PAI Item #123



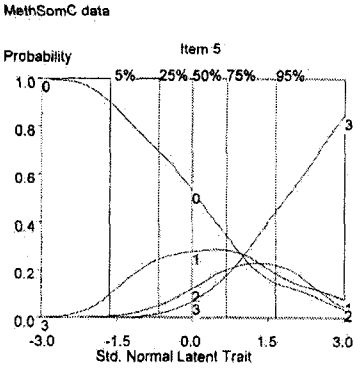


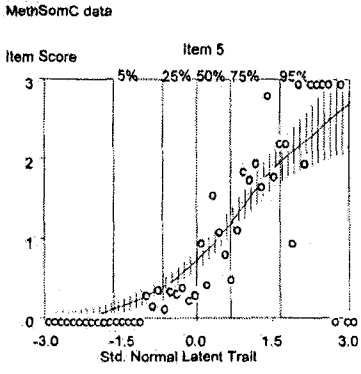
PAI Item #163



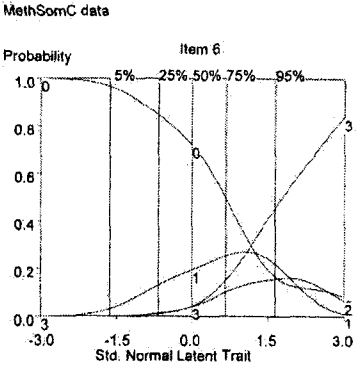


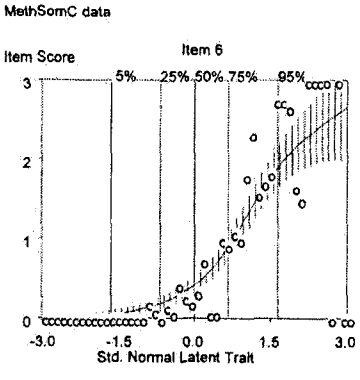
PAI Item #203



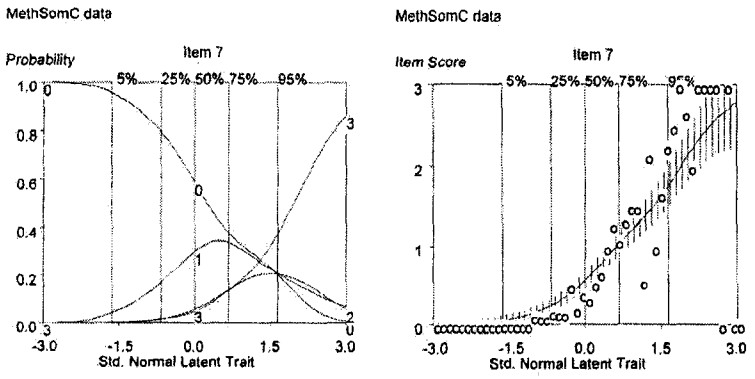


PAI Item #243





PAI Item #283



PAI Item #323

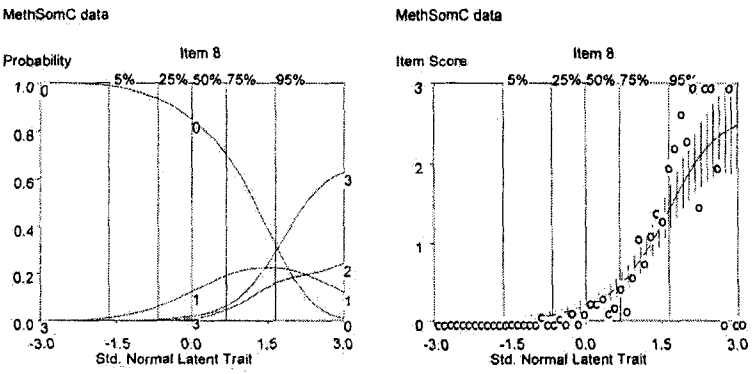


Figure 99. Option characteristic curves (OCC) and item characteristic curves (ICC) for the each item of the SOM-C scale of the PAI for the methadone maintenance treatment population (N = 323).

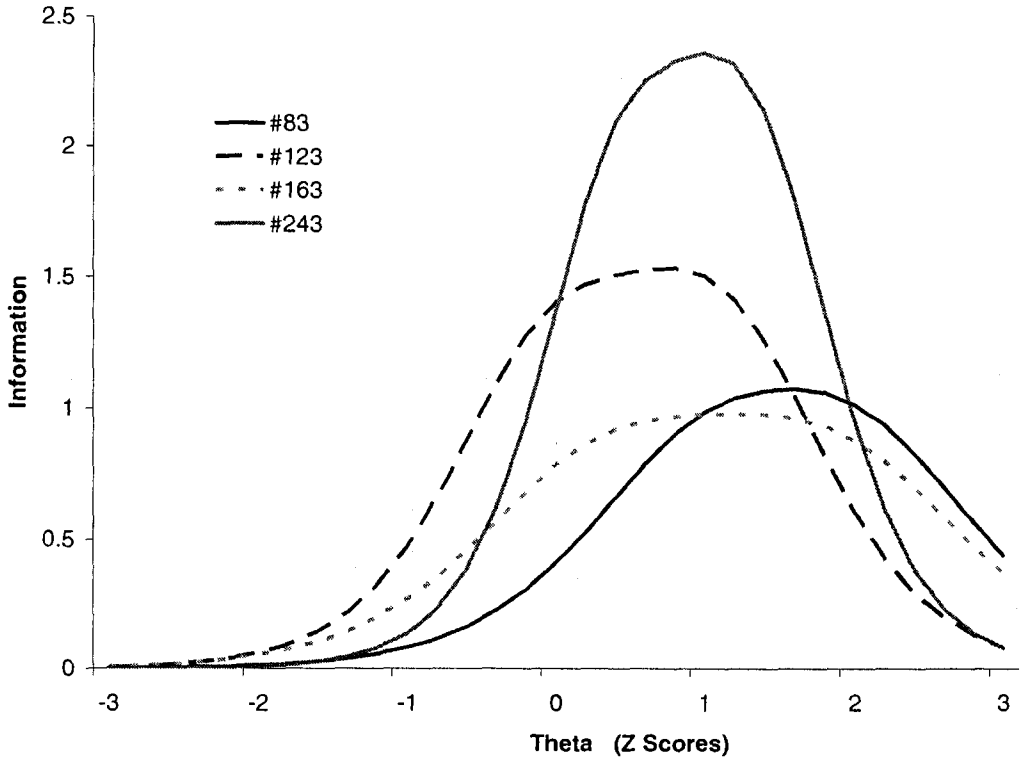


Figure 100. Item information curves for the four items of the SOM-C-R subscale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

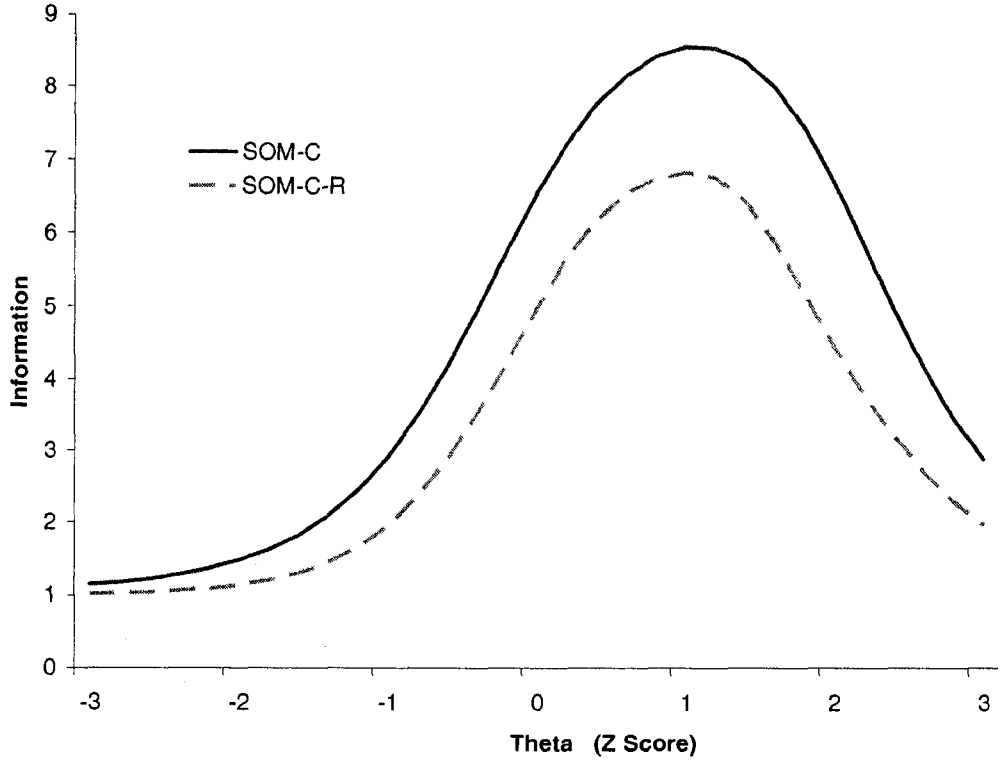
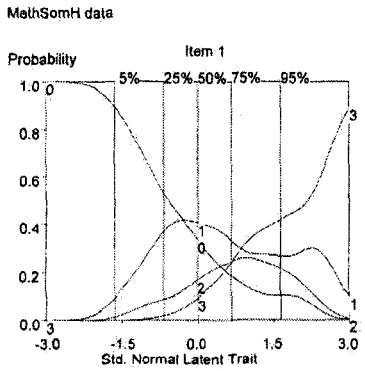
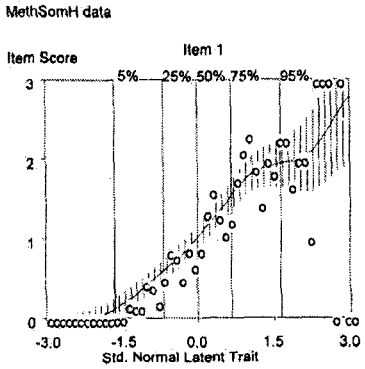


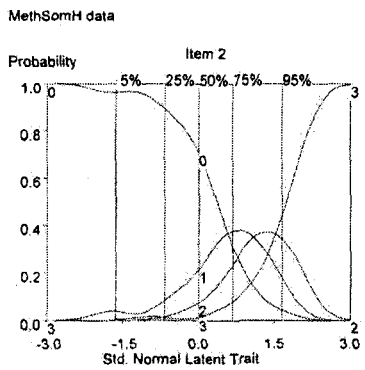
Figure 101. Comparison of total test information for the original SOM-C and the SOM-C-R based upon responses to the PAI by a sample of MMT patients (N = 323).

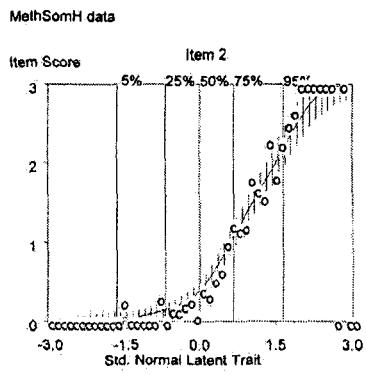
PAI Item #12



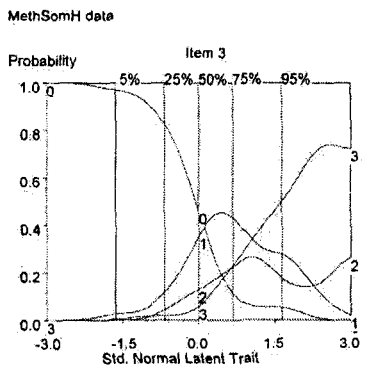


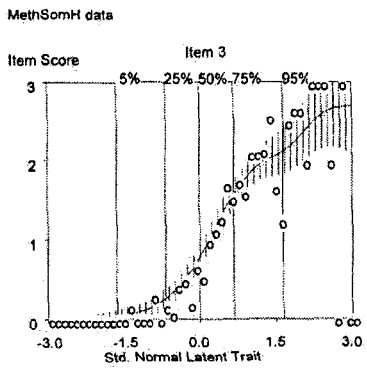
PAI Item #52



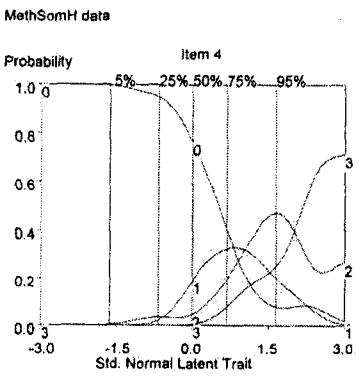


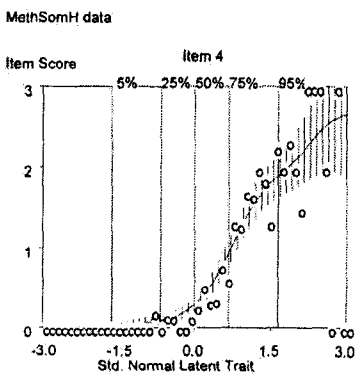
PAI Item #92



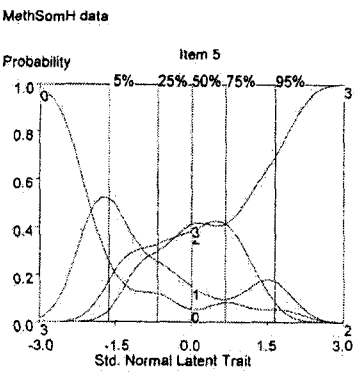


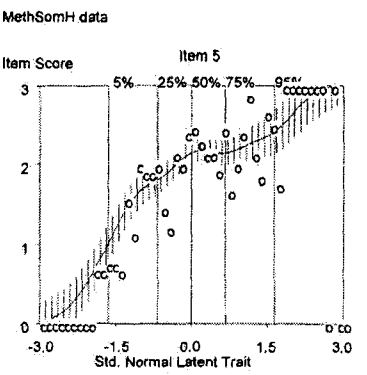
PAI Item #132



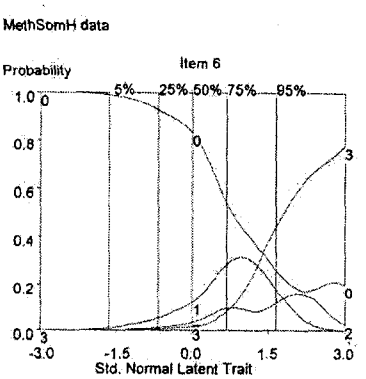


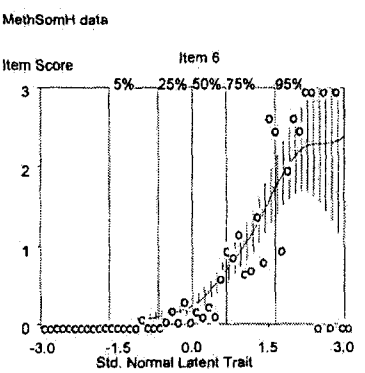
PAI Item #172



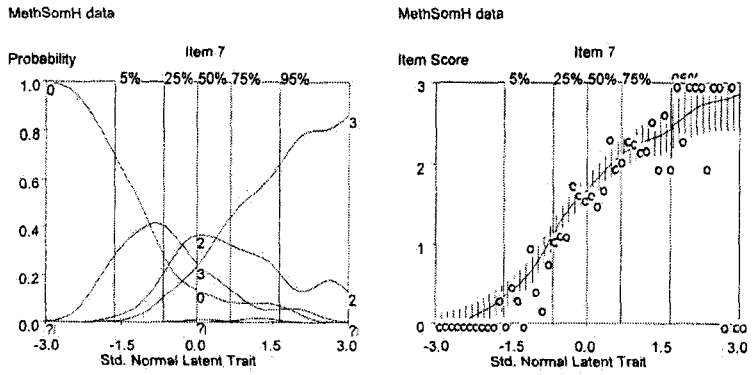


PAI Item #212





PAI Item #252



PAI Item #292

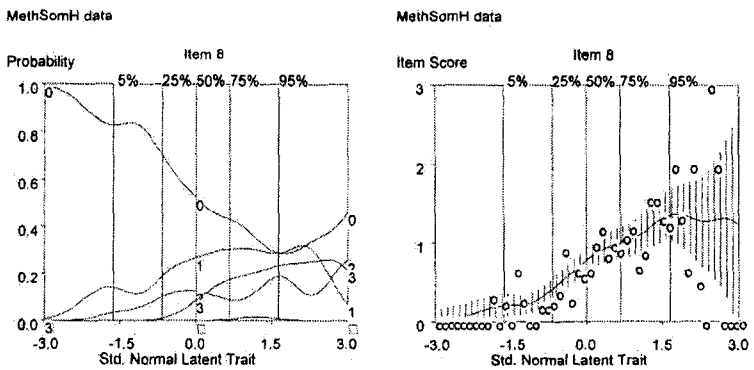


Figure 102. Option characteristic curves and item characteristic curves of the original SOM-H subscale of the PAI in use with the MMT population (N = 323).

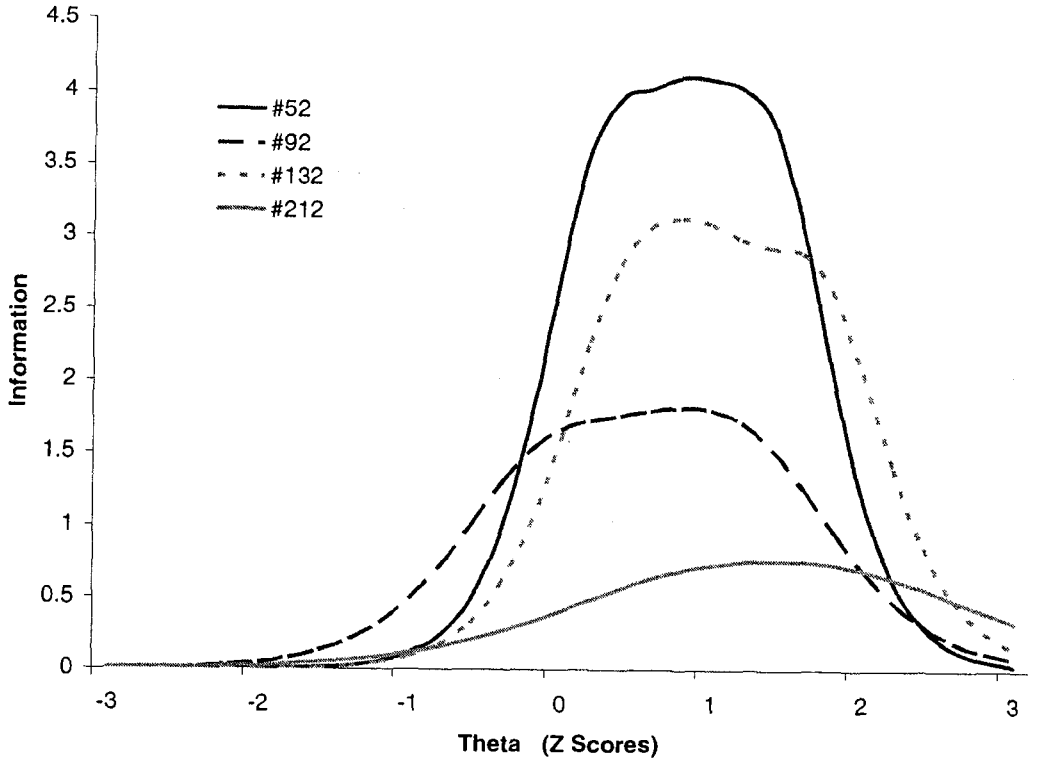


Figure 103. Item information curves for the SOM-H-R subscale items based upon responses to the PAI by a sample of MMT patients ($N = 323$).

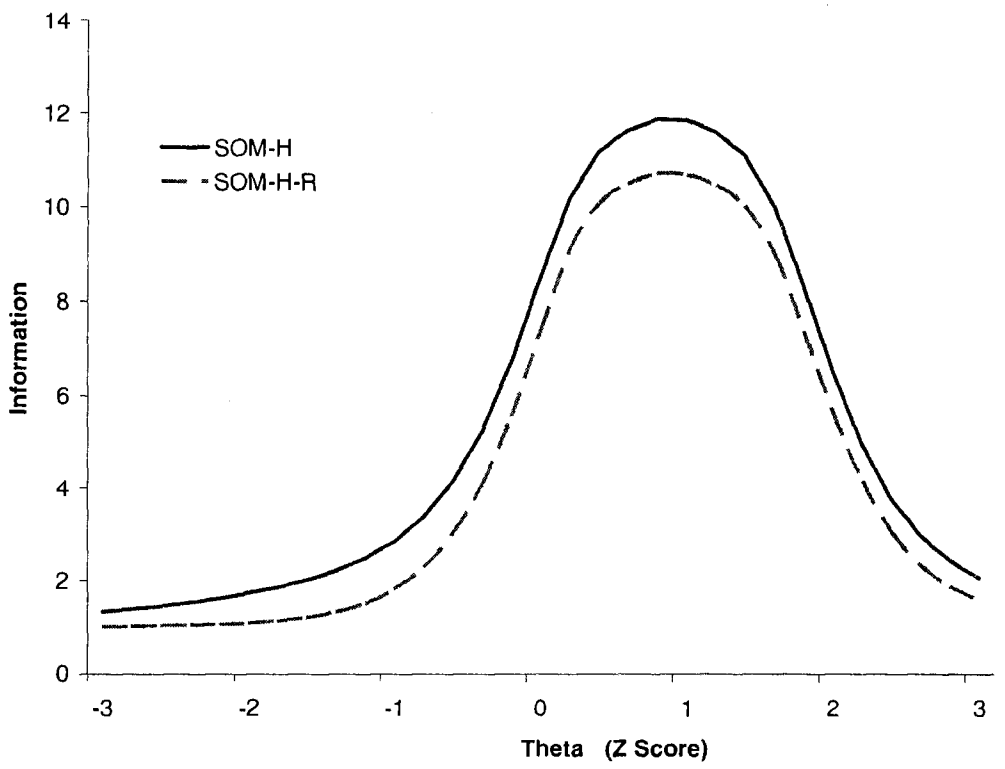
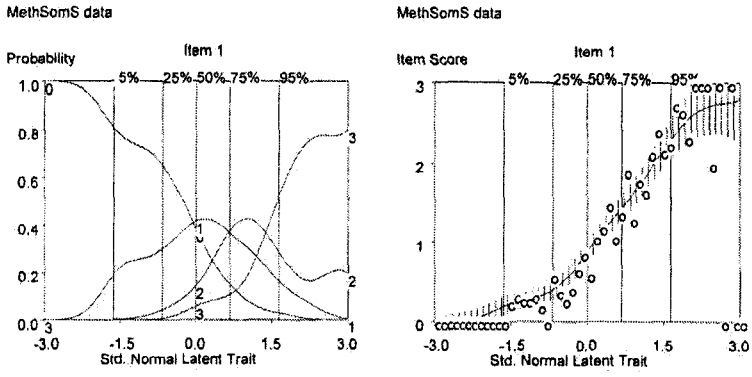
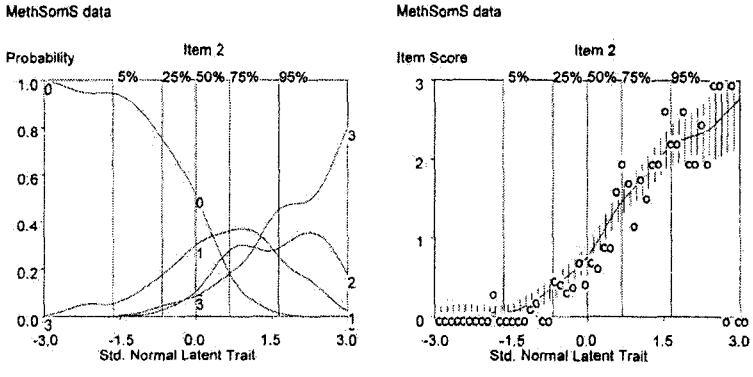


Figure 104. Comparison of total test information between the original SOM-H (8 items) subscale and the new SOM-H-R (4 items) subscale when used in the MMT population (*N* = 323).

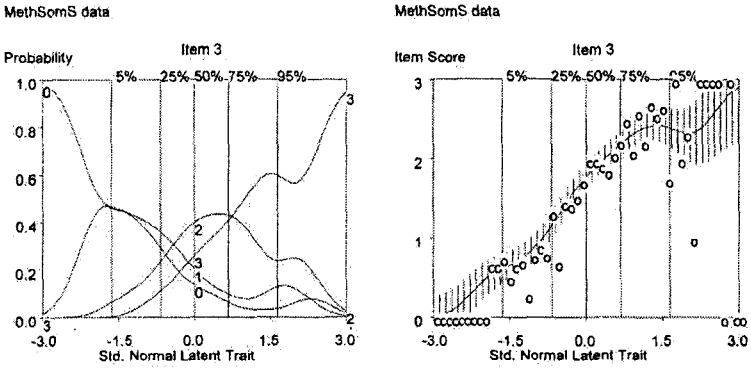
PAI Item #32



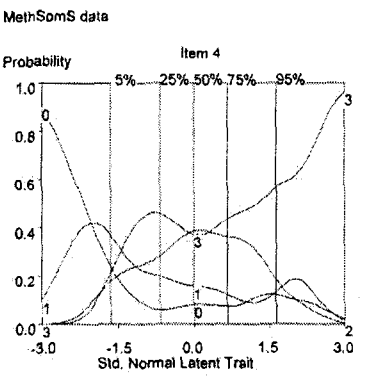
PAI Item #72

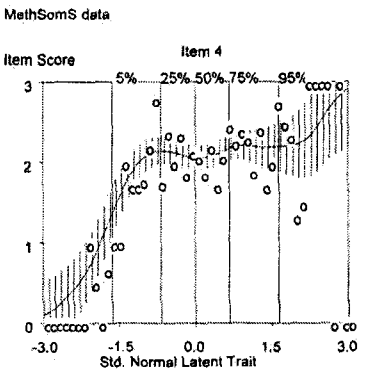


PAI Item #112

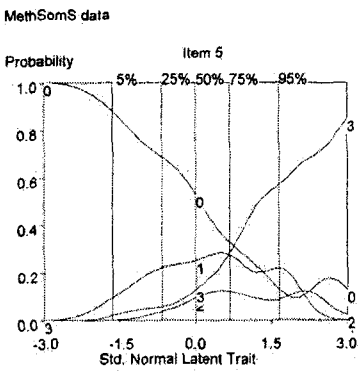


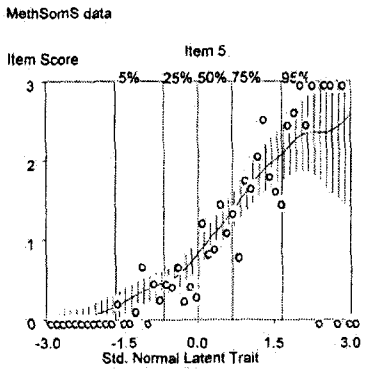
PAI Item #152



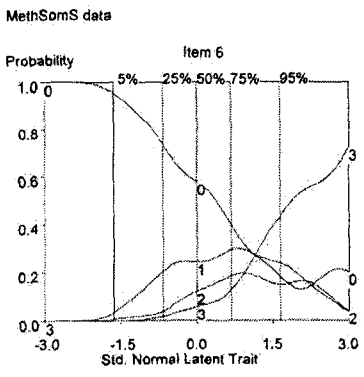


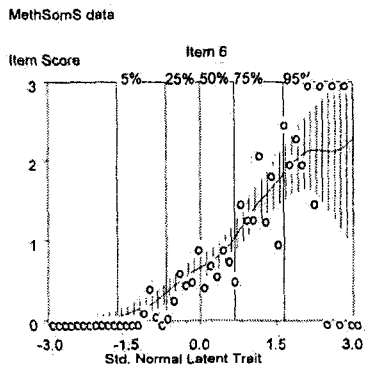
PAI Item #192



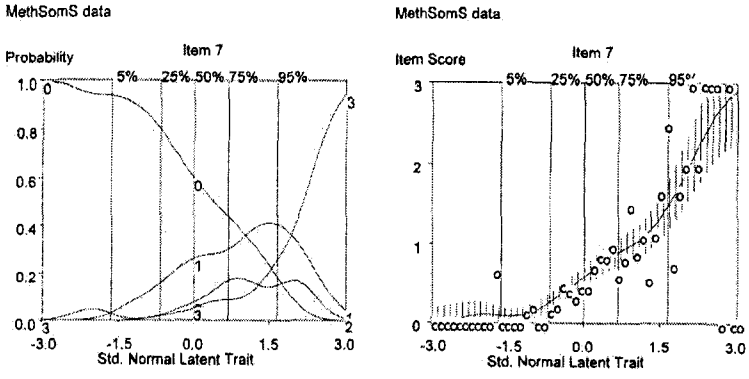


PAI Item #232





PAI Item #272



PAI Item #312

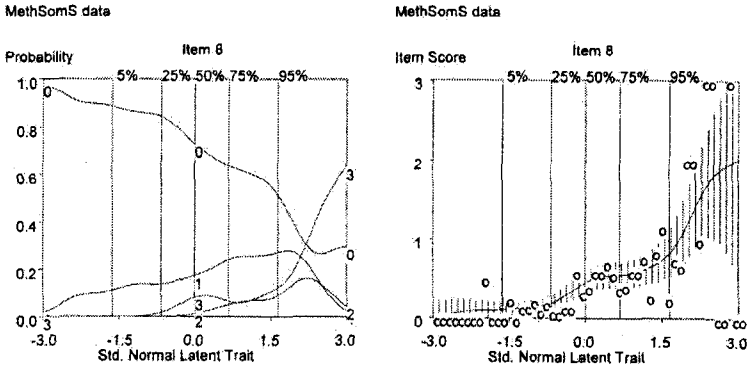


Figure 105. Option and Item characteristic curves of the SOM-S subscale when employed in a sample of MTT patients (N = 332).

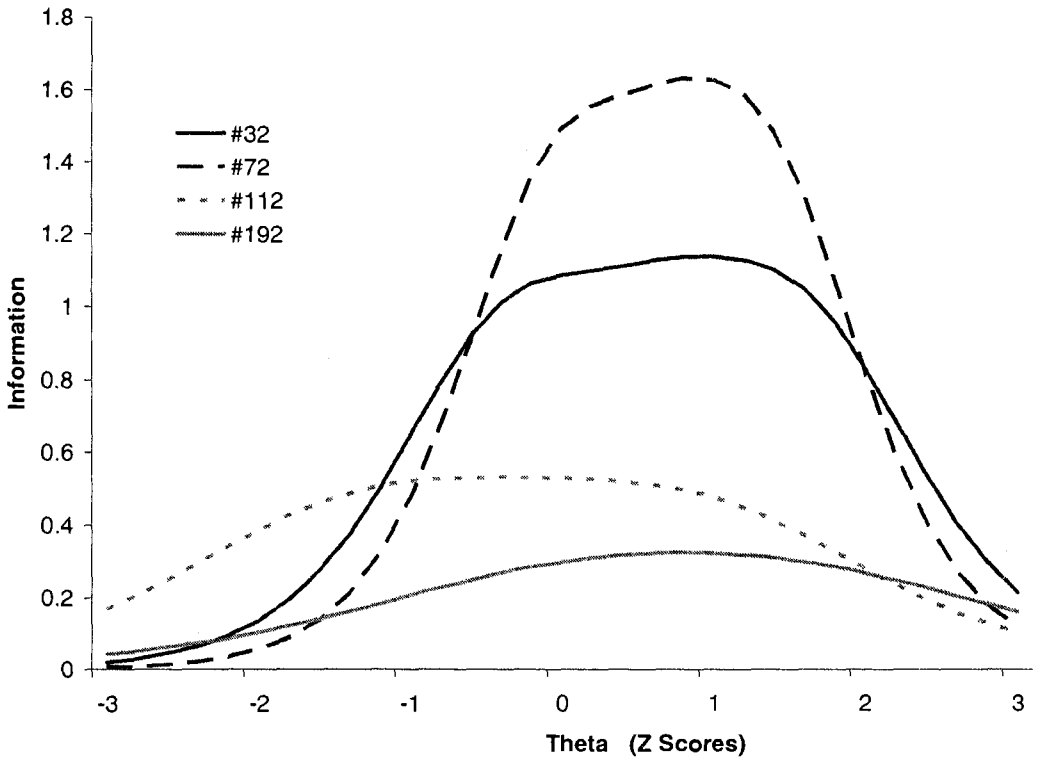


Figure 106. Item information curves for the SOM-S-R subscale items based upon responses to the PAI from a sample of MMT patients (N = 323).

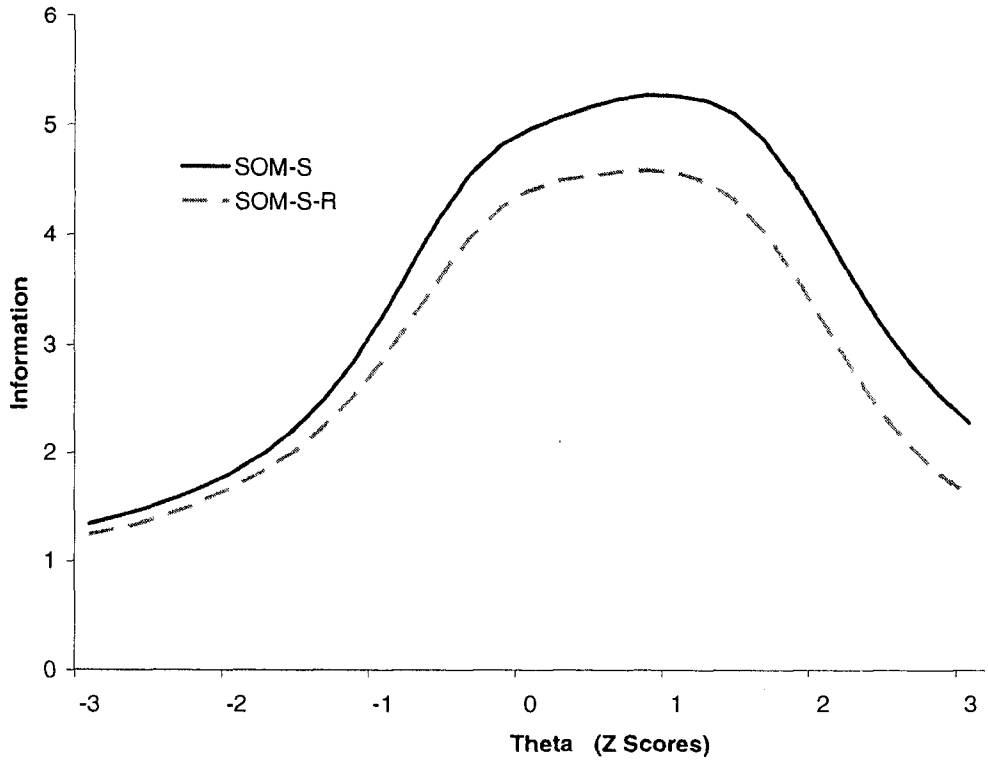
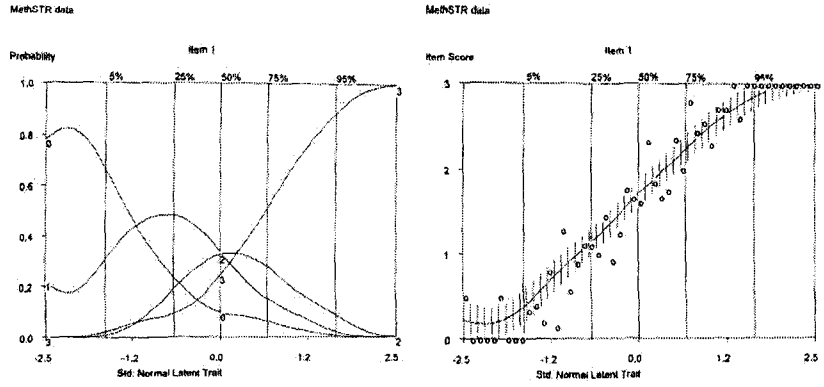
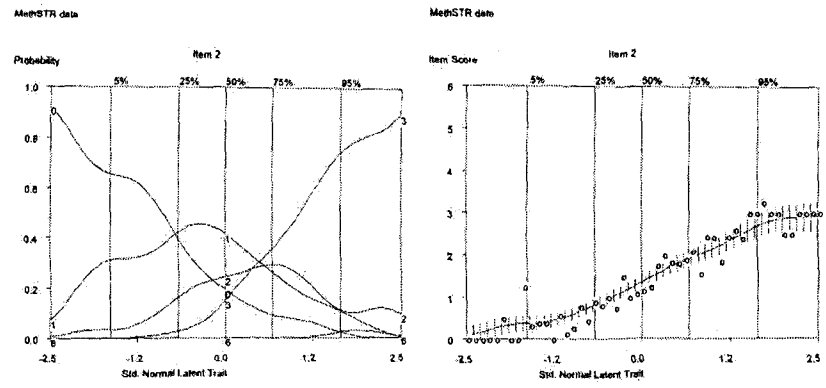


Figure 107. Comparison of total test information between the original SOM-S (8 items) subscale and the new SOM-S-R (4 items) subscale when used in the MMT population (*N* = 323).

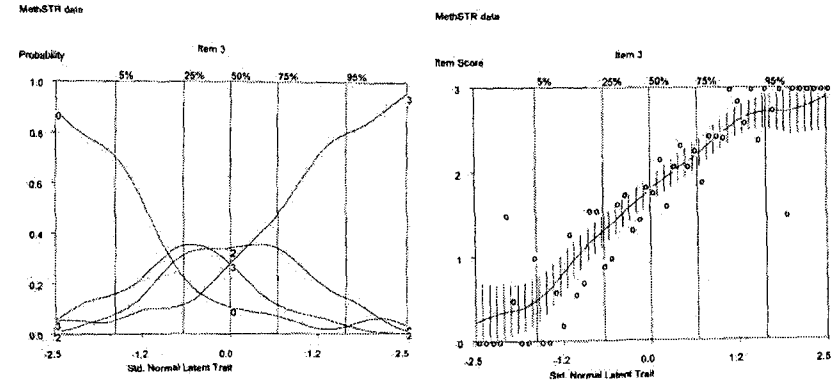
PAI Item #321



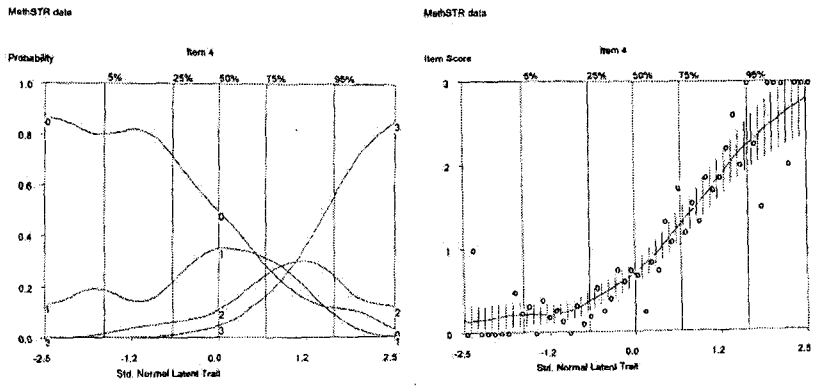
PAI Item #322



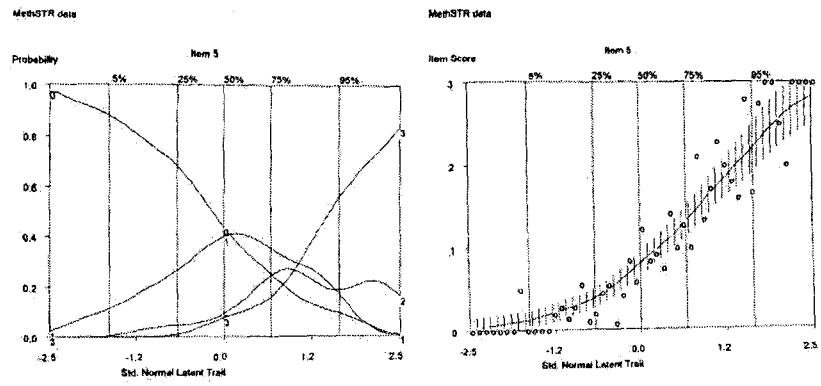
PAI Item #323



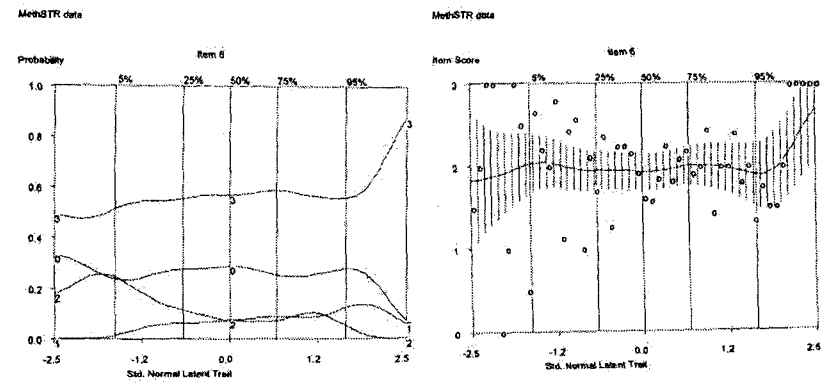
PAI Item #324



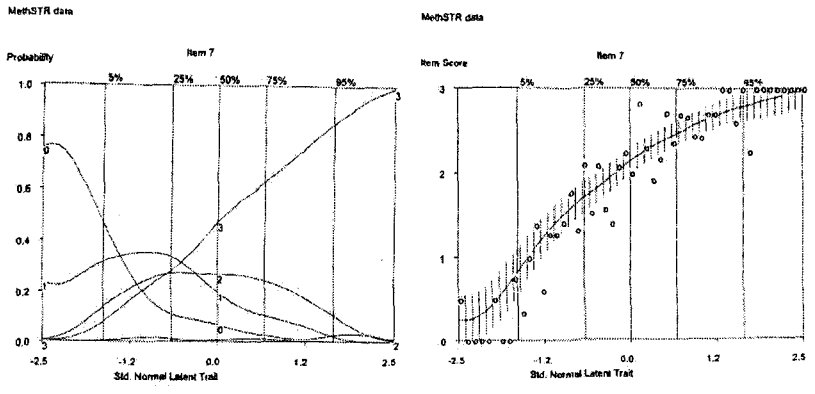
PAI Item #325



PAI Item #326



PAI Item #327



PAI Item #328

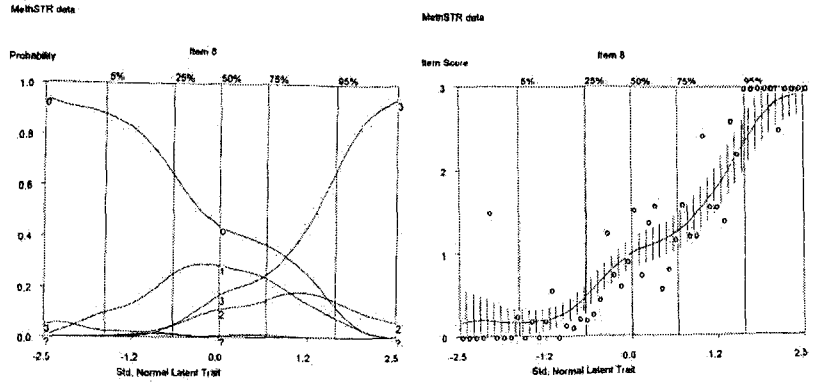


Figure 108. Option and item characteristic curves for the PAI scale STR based upon the responses to the PAI by a sample of MMT patients (N = 323).

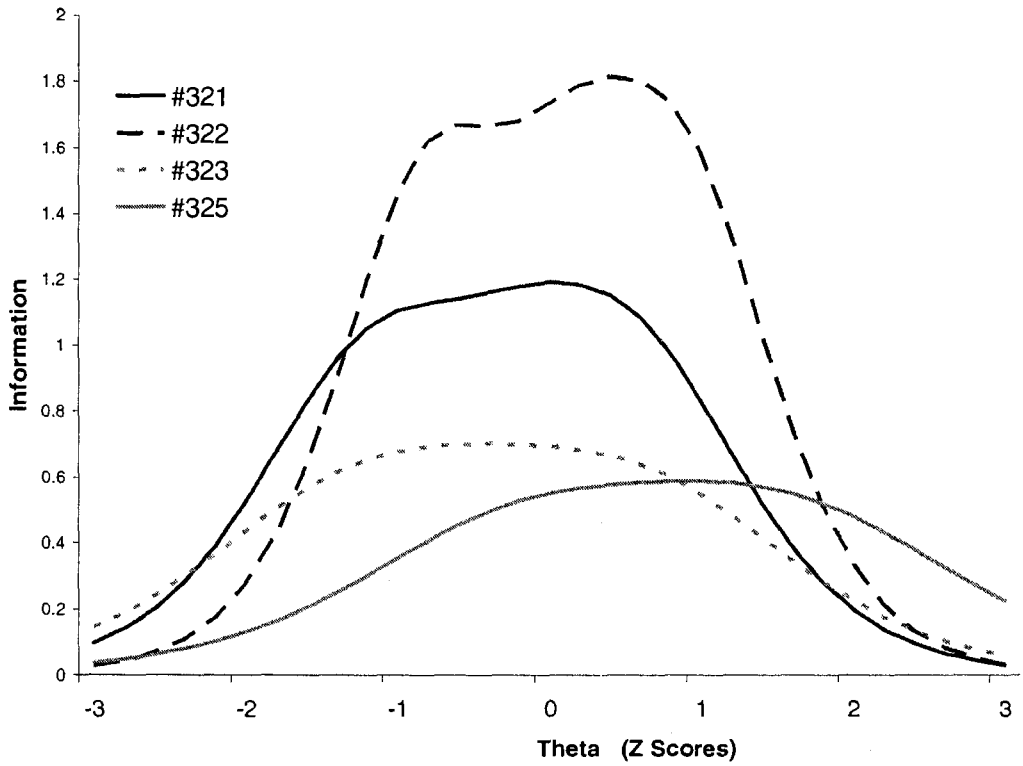


Figure 109. Item information curves for the STR-R scale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

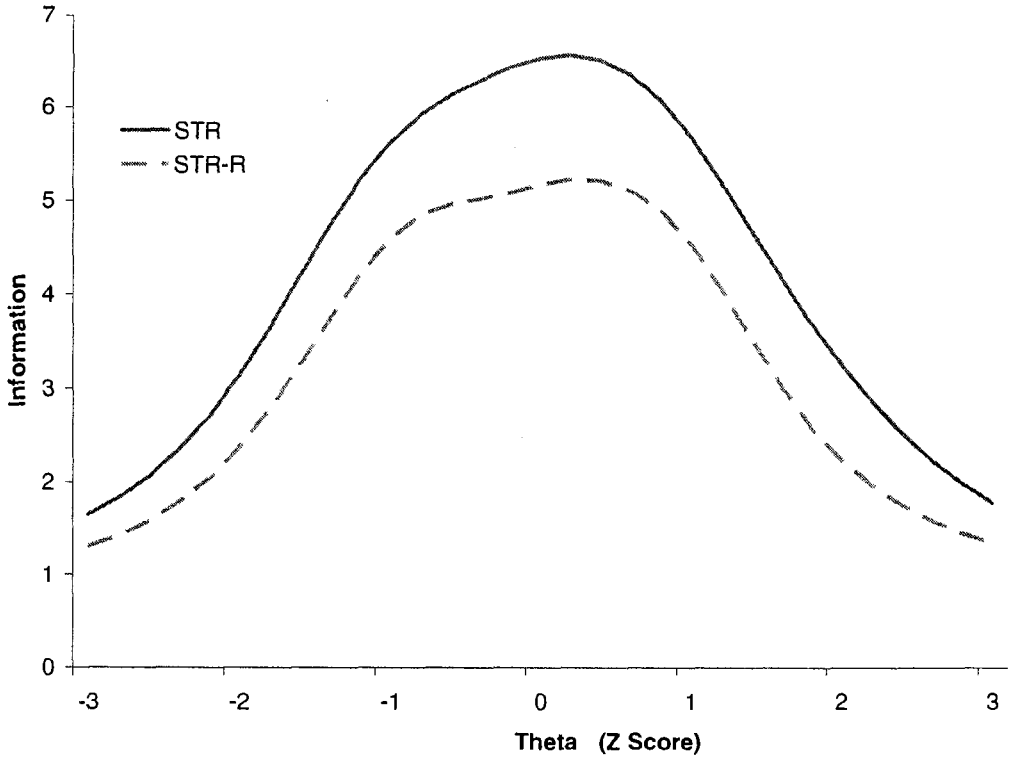
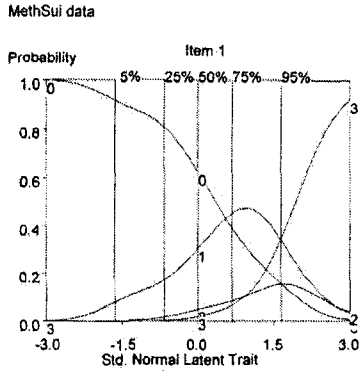
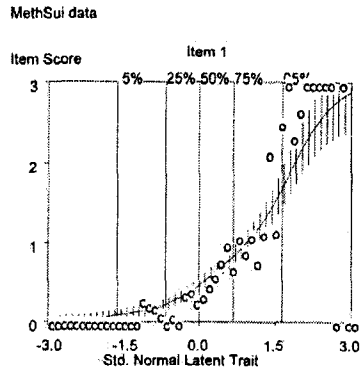


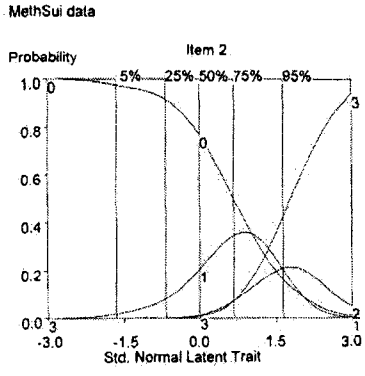
Figure 110. Total scale information curves comparing the original STR scale with the STR-R scale based upon responses to the PAI by a sample of MMT patients ($N = 323$).

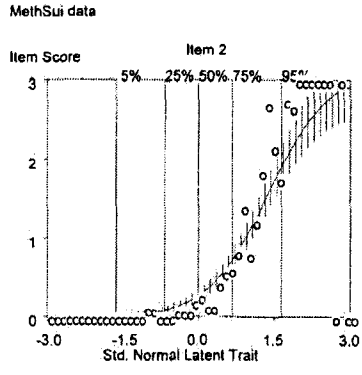
PAI Item #20



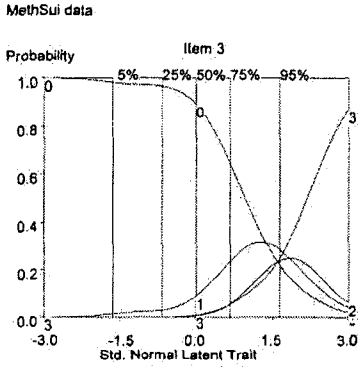


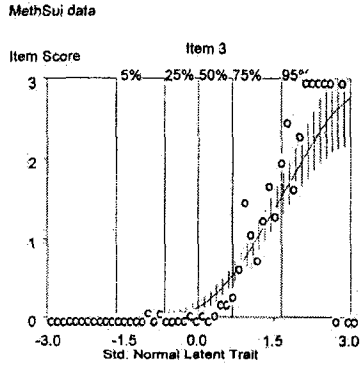
PAI Item #60



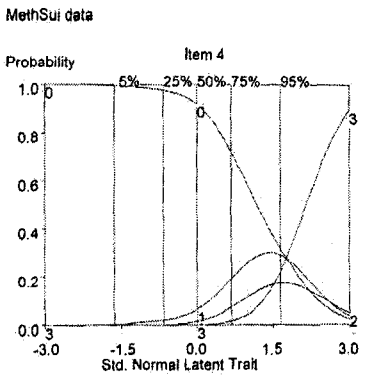


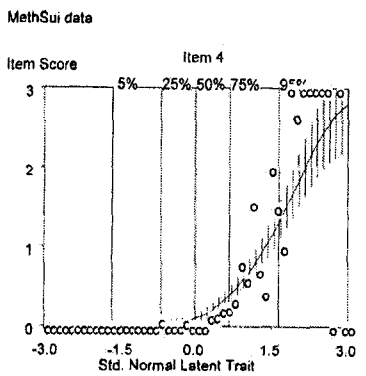
PAI Item #100



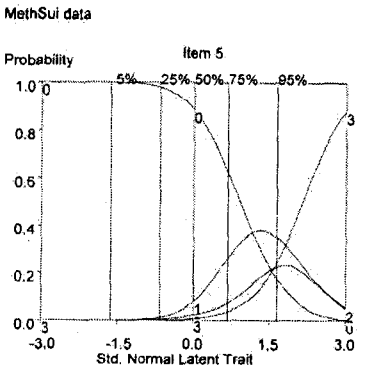


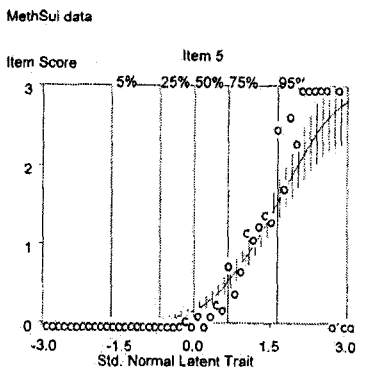
PAI Item #140



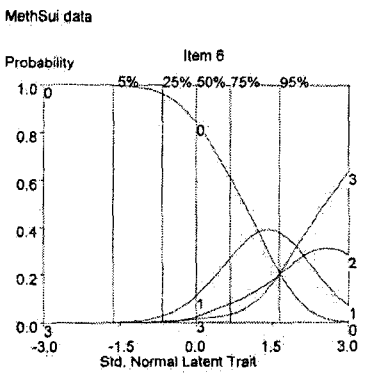


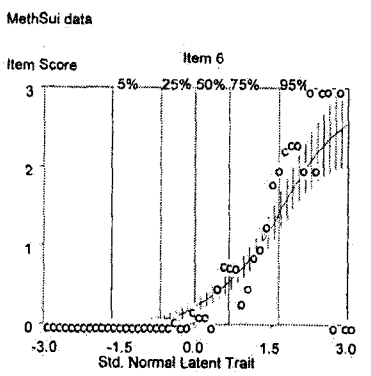
PAI Item #180



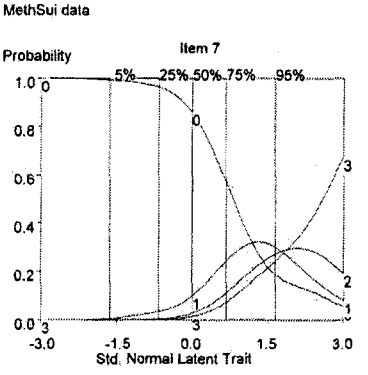


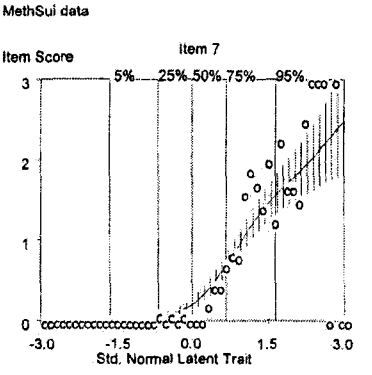
PAI Item #220



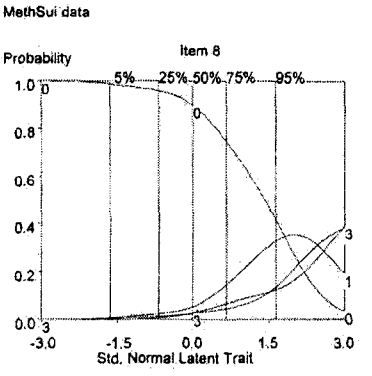


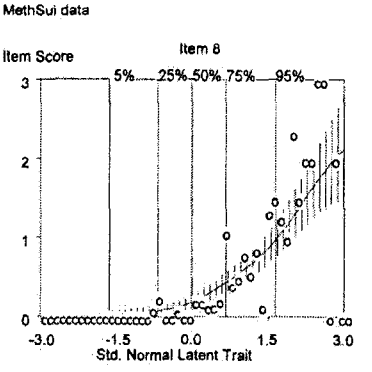
PAI Item #260



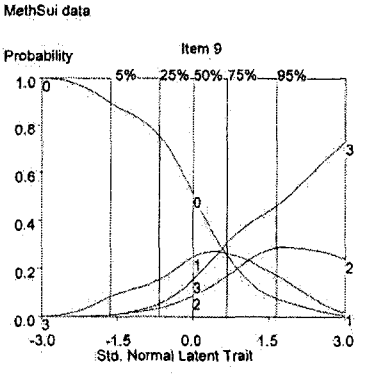


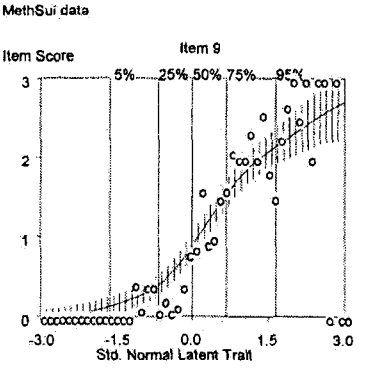
PAI Item #261



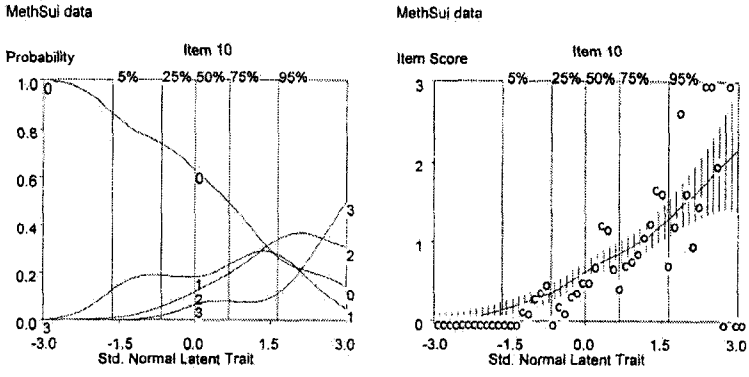


PAI Item #300

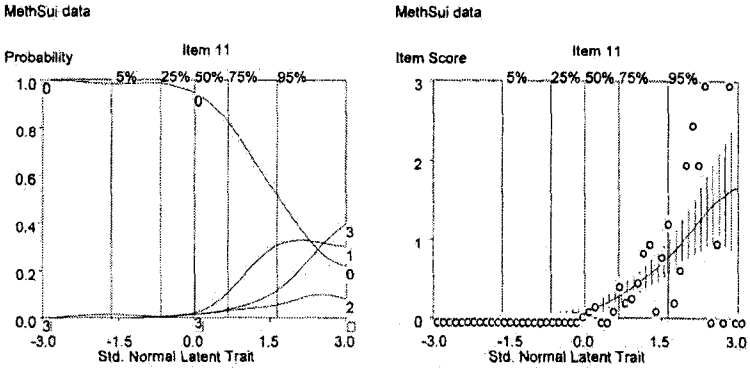




PAI Item #301



PAI Item #340



PAI Item #341

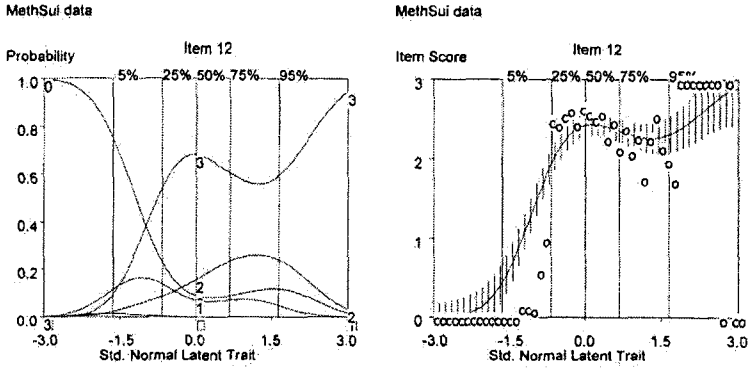


Figure 111. Option characteristic curves and item characteristic curves for each item of the SUI scale based upon the responses of a sample of MMT patients (N = 323).

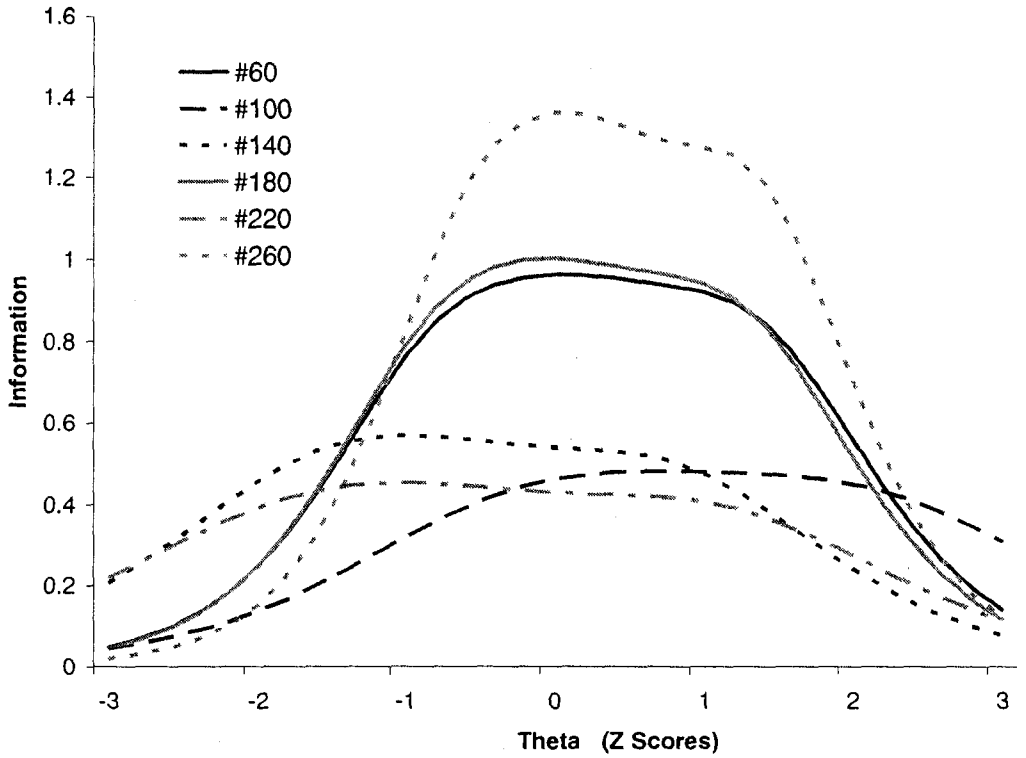


Figure 112. Option and item information curves for the final set of items for the SUI-R scale in use with a MMT population ($N = 323$).

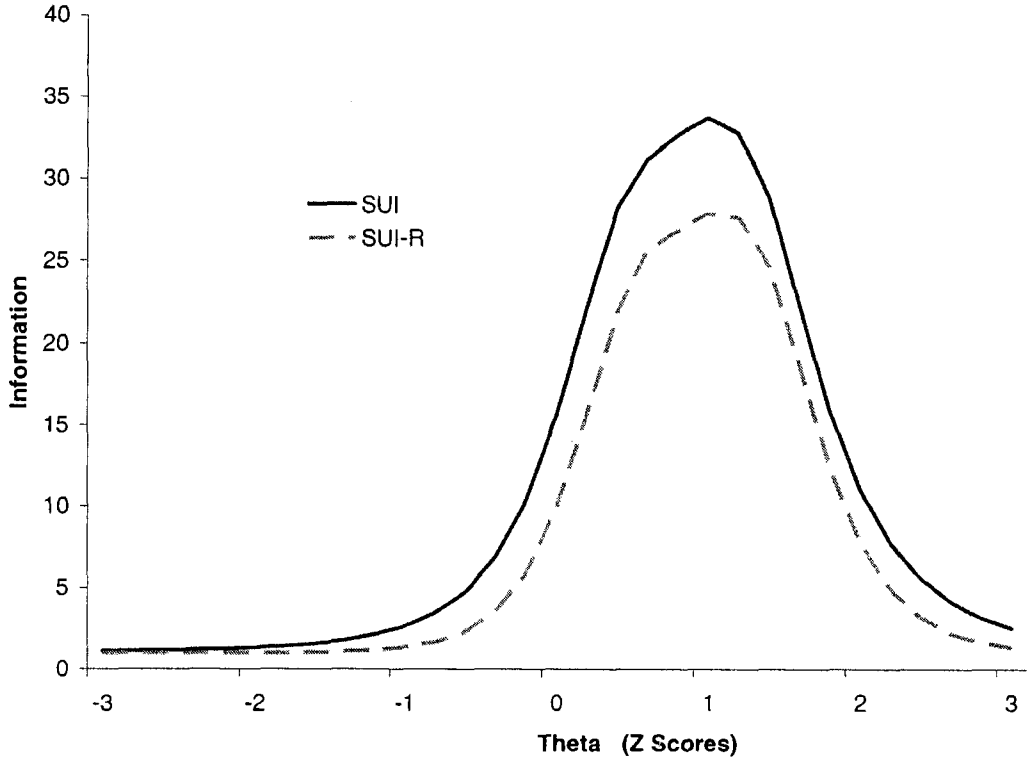
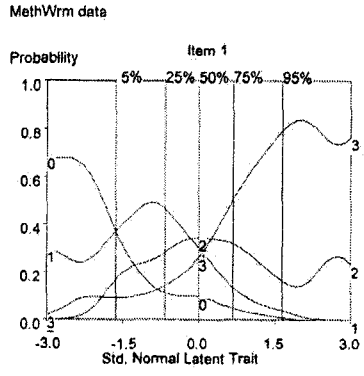
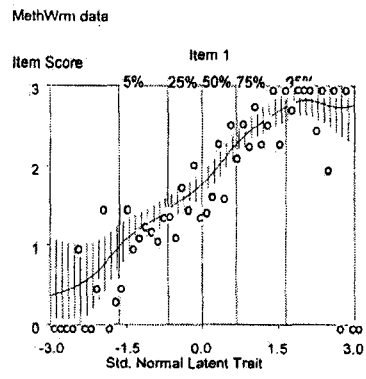


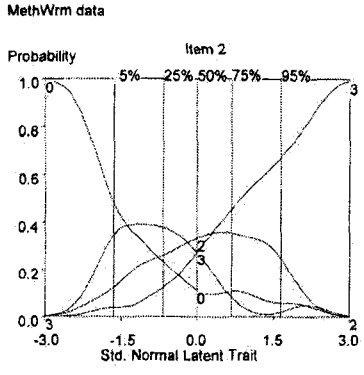
Figure 113. Total information curves for the SUI-R (5 items) versus initial estimates for the original SUI (9 items) based upon responses to the PAI from a sample of MMT patients ($N = 323$).

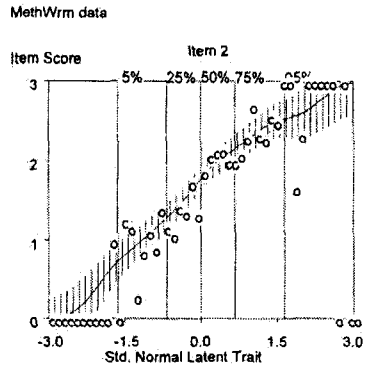
PAI Item #13



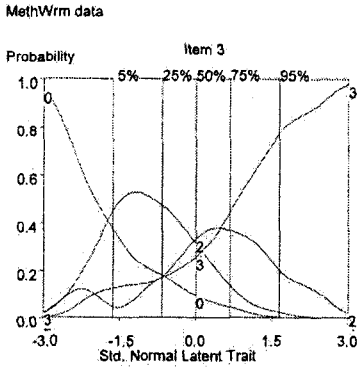


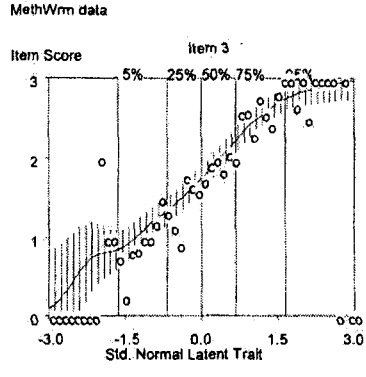
PAI Item #53



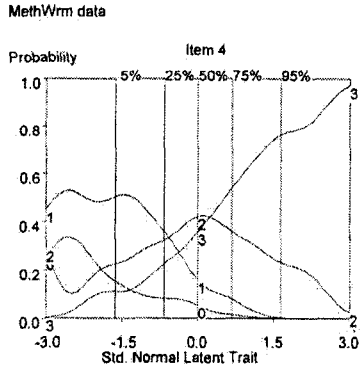


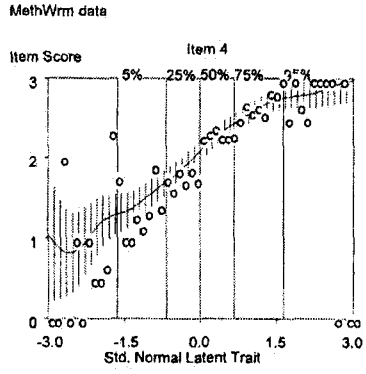
PAI Item #93



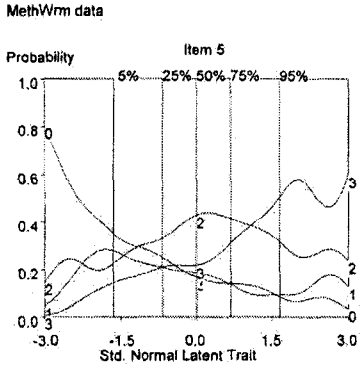


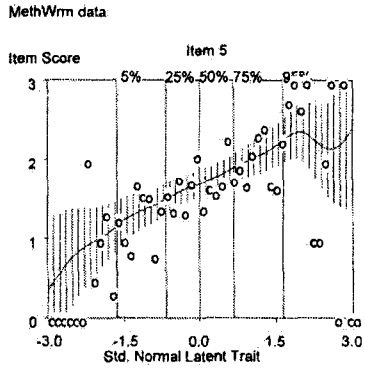
PAI Item #133



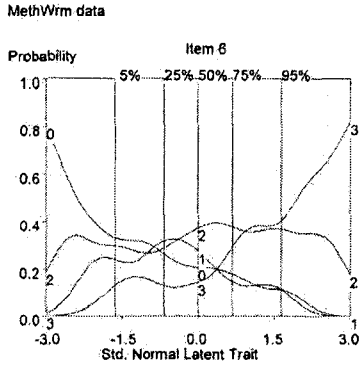


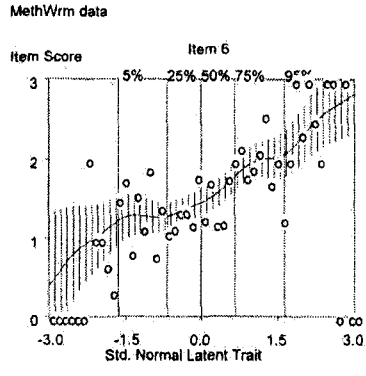
PAI Item #173



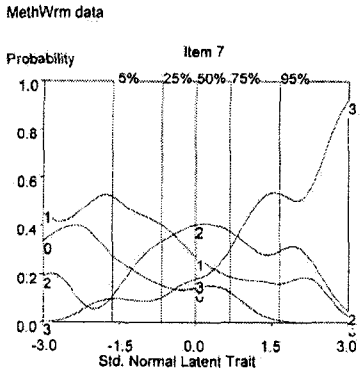


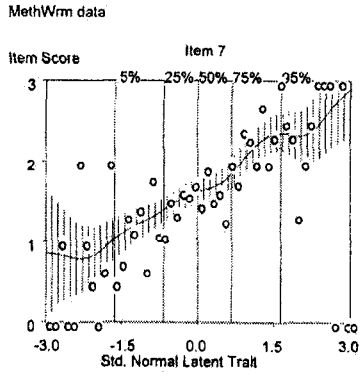
PAI Item #213



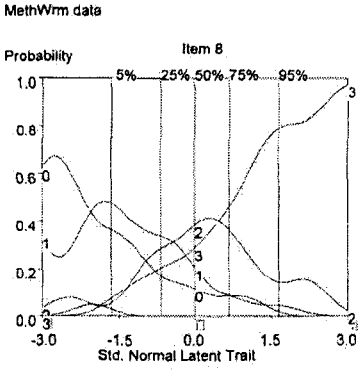


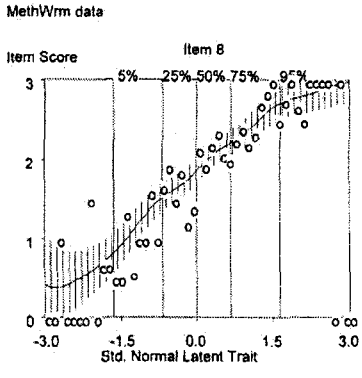
PAI Item #253



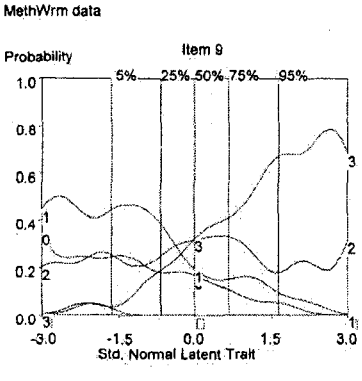


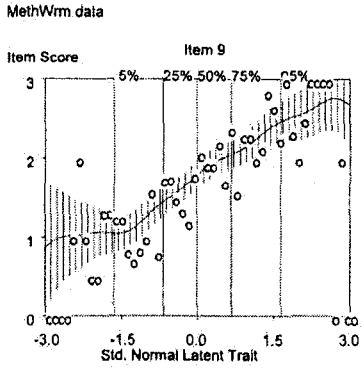
PAI Item #293



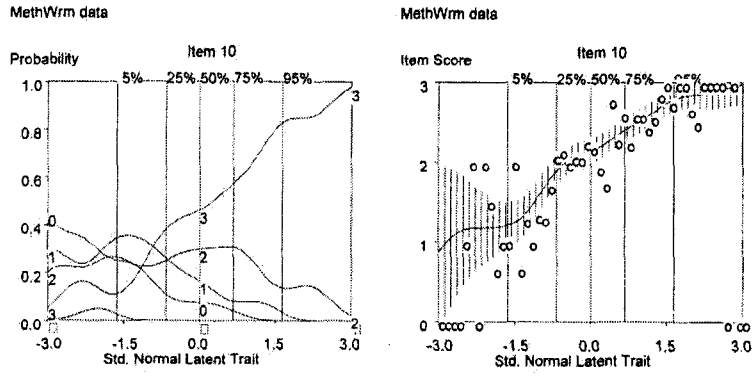


PAI Item #330

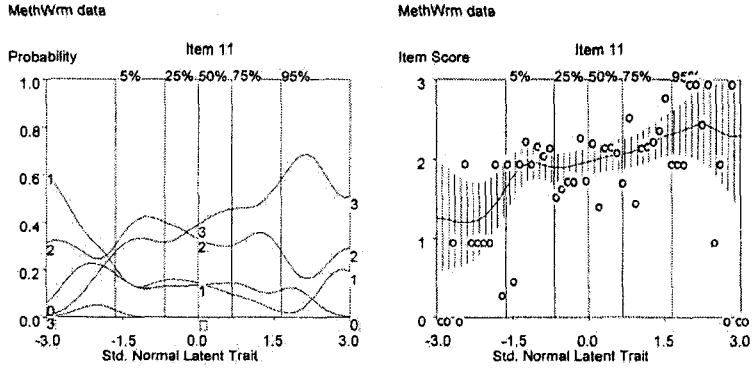




PAI Item #331



PAI Item #332



PAI Item #333

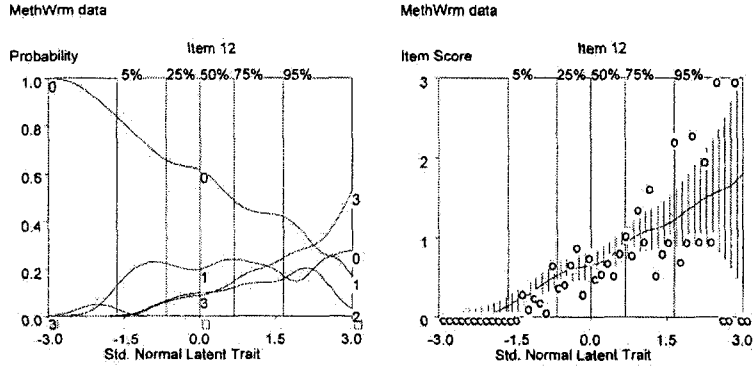


Figure 114. Option and item characteristic curves for the items of the WRM scale based upon responses to the PAI in a sample of MMT patients ($N = 323$).

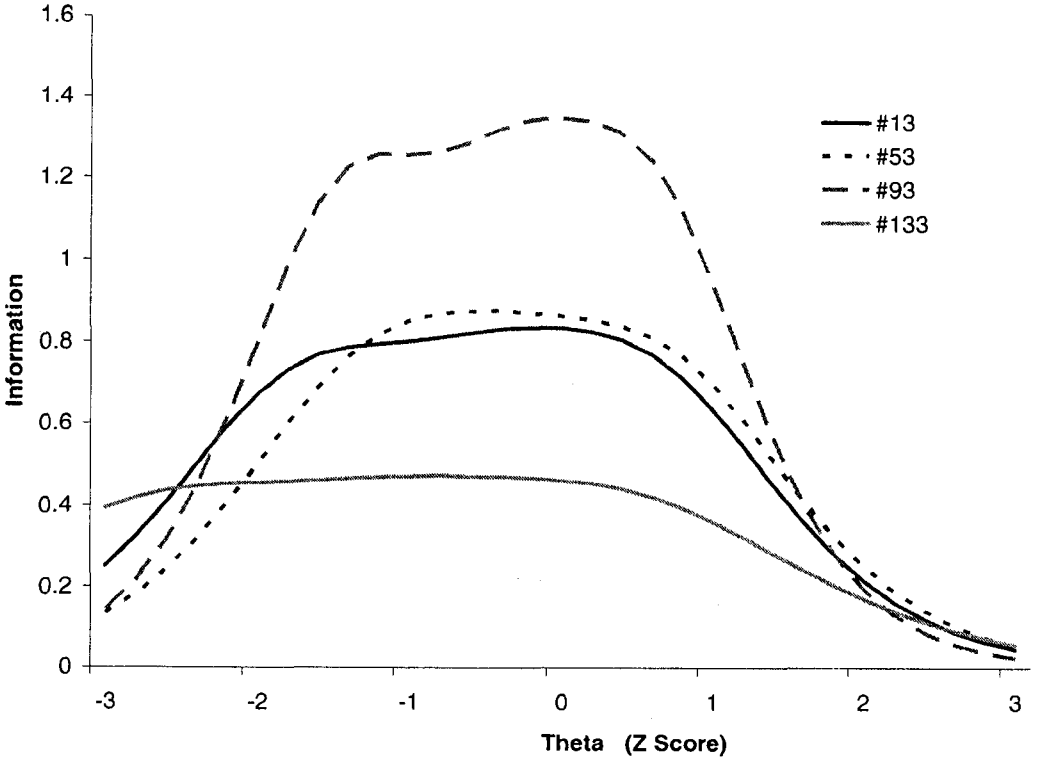


Figure 115. Item information curves for the WRM-R scale based upon responses to the PAI in a sample of MMT patients ($N = 323$).

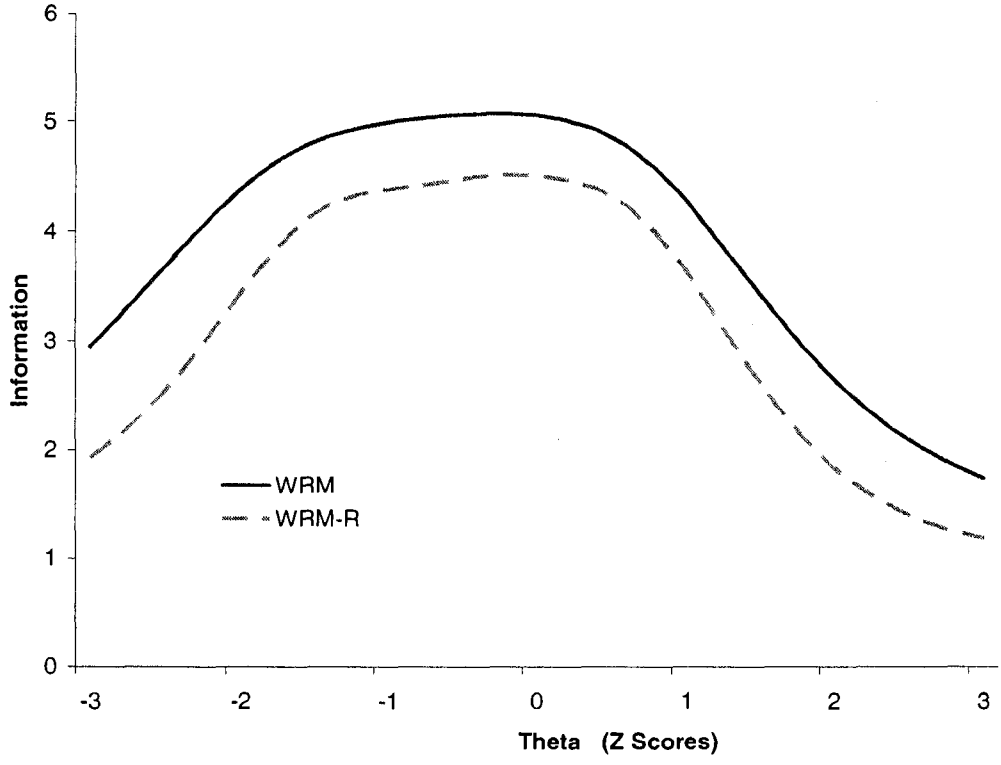
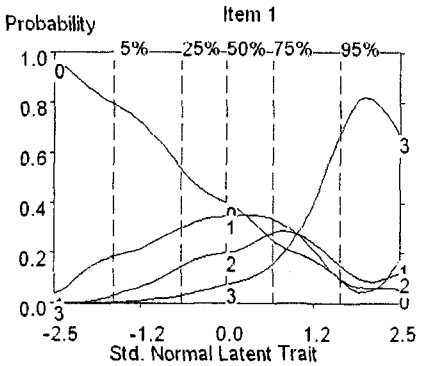
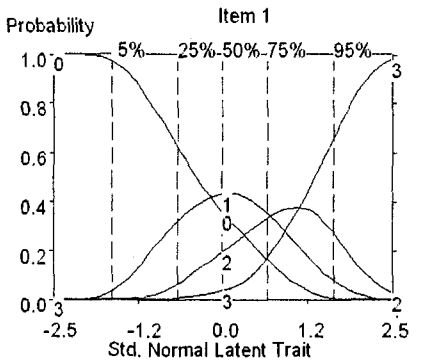


Figure 116. Total item information curves comparing the original 12 item WRM scale and the four item WRM-R scale based upon responses to PAI by a sample of MMT patients ($N = 323$).

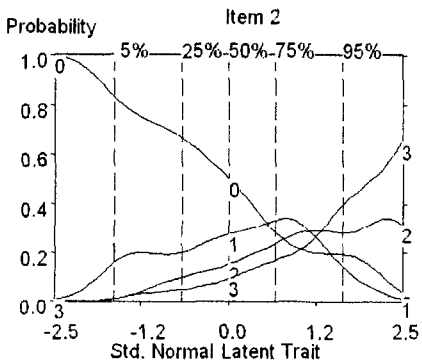
PAI Item #5 (ARD-O)



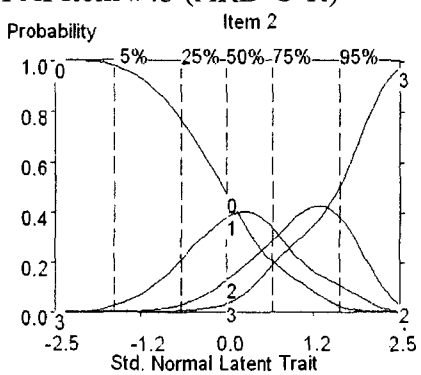
PAI Item #5 (ARD-O-R)



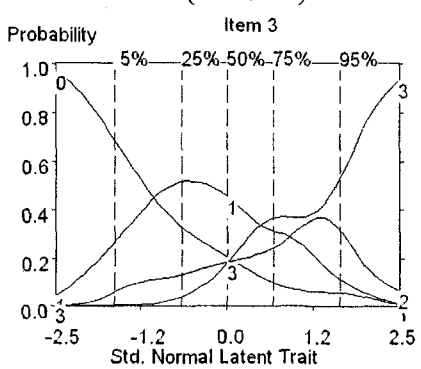
PAI Item #45 (ARD-O)



PAI Item #45 (ARD-O-R)



PAI Item #85 (ARD-O)



PAI Item #85 (ARD-O-R)

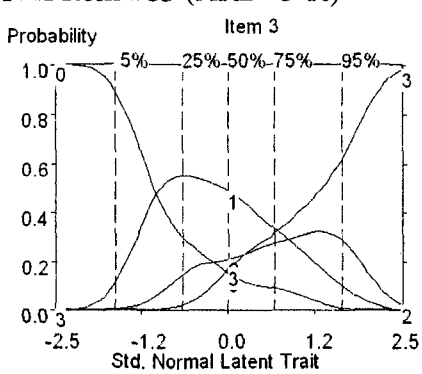


Figure 117. Comparison of OCC for the first three items of the original eight item ARD-O subscale, compared to curves for same items calibrated following the removal of the five items which demonstrated poor psychometric function based upon responses to the PAI from a sample of MMT patients (N = 323).

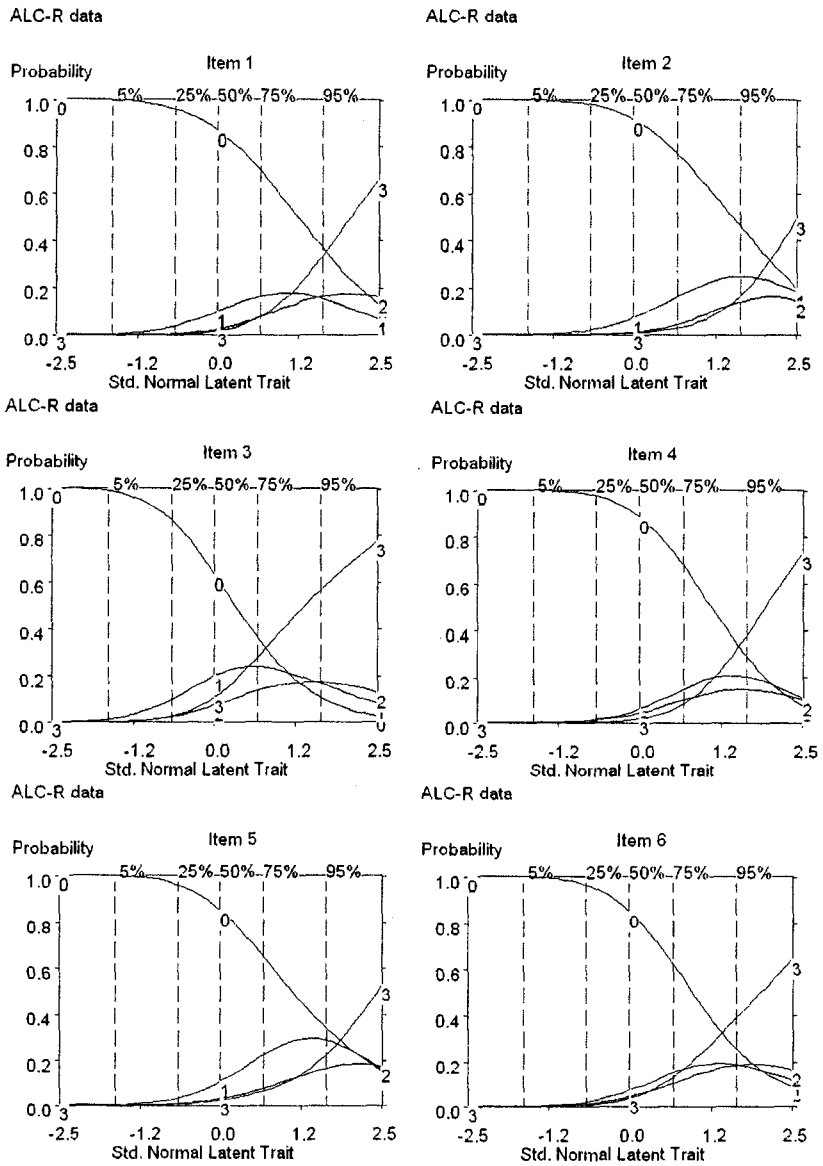


Figure 118. Option characteristic curves for the six items of the ALC-R subscale plotted using Gaussian kernel smoothing techniques indicating that this scale would likely benefit from an altered scoring algorithm based upon responses to the PAI from a sample of MMT patients (N = 323).