A CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY ANALYSIS OF THE DSM-IV SYMPTOM CRITERIA FOR A MAJOR DEPRESSIVE EPISODE USING DATA FROM THE NATIONAL COMORBIDITY SURVEY – REPLICATION

by

ANTHONY P. PAWLAK

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

written under the direction of

Douglas A. Penfield, Ph.D.

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2010

ABSTRACT OF THE DISSERTATION

A Classical Test Theory and Item Response Theory Analysis of the DSM-IV Symptom Criteria for a Major Depressive Episode Using Data from the National Comorbidity Survey – Replication

By ANTHONY P. PAWLAK

Dissertation Director:
Douglas A. Penfield, Ph.D.

Formal psychiatric symptom criteria are used to delineate the boundary between "normal" and "abnormal" behavior. In North America, the current official psychodiagnostic criteria for a multitude of psychiatric disorders are codified in the *Diagnostic and Statistical Manual of Mental Disorders* (4th Edition, text revision) (APA, 2000). Psychodiagnostic symptom criteria are indicators of psychopathological constructs that are clearly latent, however, it is somewhat astonishing that formal psychometric techniques that have been developed to model latent constructs have not been used to develop and evaluate psychodiagnostic symptom criteria (Aggen, Neale, & Kendler, 2005; Zimmerman, McGlinchey, Young, & Chelminski, 2006a, 2006b).

There are two main psychometric paradigms that are currently in use: classical test theory and item response theory (Crocker & Algina, 1986). Classical test theory has been extensively used on both cognitive constructs and noncognitive constructs (Crocker & Algina, 1986; Embretson & Hershberger, 1999). Item response theory is considered to be theoretically superior to classical test theory and it has revolutionized the creation and evaluation of cognitive constructs (Crocker & Algina, 1986; Embretson & Hershberger, 1999; McDonald, 1999). However, item response theory has not been extensively

utilized for the creation and evaluation of noncognitive constructs, even though it holds great promise in this regard (Reise, 1999; Reise & Henson, 2003).

The proposed study will use classical test theory and item response theory to assess the psychodiagnostic symptom criteria for depression as found in the *Diagnostic and Statistical Manual of Mental Disorders* (4th Edition, text revision) (APA, 2000). The data to be used in the proposed study was collected in the National Comorbidity Survey – Replication, which was a nationally representative epidemiological community survey (Kessler et al., 2004; Kessler & Merikangas, 2004). The results of such a study will give a sophisticated psychometric perspective on the psychodiagnostic symptom criteria of depression that has not yet been available and it will provide valuable information on improving and refining future diagnostic symptom criteria of depression.

Acknowledgements

First, I would like to thank my family (my parents and my sister Krystina) for their continual support throughout my tenure in graduate school. This doctoral dissertation would not have been possible without them!

I would like to thank Douglas A. Penfield, PhD for guiding me through the process of writing the dissertation and helping me navigate some of the more bureaucratic aspects of putting together a doctoral dissertation, Gregory Camilli, PhD for invaluable advice on the doctoral dissertation project's data analysis and its interpretation, Jimmy de la Torre, PhD for teaching me advanced psychometric theory, especially item response theory, and Jim Langenbucher for initially helping give birth to the idea behind this dissertation's research project and providing guidance in determining the practical interpretations of my results.

Two of my professors from my undergraduate studies at Franklin and Marshall College, Michael Penn, PhD and Thomas Hopkins, PhD, provided much moral support to stay the course and complete my doctoral dissertation. I would also like to thank George Rosenstein, PhD, also from Franklin and Marshall College, for teaching me mathematics and how to apply it to real-world data, which proved essential to my graduate studies of applied statistics, psychometrics, and social research methodology.

While conducting the research for the dissertation and writing up its results, I had many friends who helped me survive the stress associated with this project. In particular, from my F&M days, I would like to thank Derek Webb, Matt Kernicky, Chuck Valentine and Jeff French for giving me much needed time and space to unwind and to escape from the rigors of graduate life. From my time at Rutgers, I would like to thank my friends

from Communion and Liberation, especially Martina Saltamacchia, Michele Monserrati, Alex-Noelle O'Brien, Sarah Strenio, and Mike Erickson; from Trinity House, especially Emily Edenfield, Barbara Heck, and John Larson; and from Canterbury House, especially Allie Graham, Tim Palmer, Sam Bassler, and Gregory Bezilla, for providing much moral and spiritual support. A special mention goes to D. Paul La Montagne, PhD, who became an intellectual sounding board for many of my thoughts and ideas at crucial junctures during the process of developing and refining my dissertation research. He provided many astute comments and criticisms concerning my research from an outside perspective.

Table of Contents

Lists of Figures

List of Tables

CHAPTER I. STATEMENT OF THE PROBLEM

History of Psychiatric Diagnostic Criteria

It appears that throughout human history, there has been a strong sense among different cultures that mental states and/or behaviors fall into "normal" and "abnormal" categories, and, as a result, the medical communities of these different cultures attempted to systematically categorize the various kinds of mental and/or behavioral abnormalities into psychiatric diagnostic systems (Frances, Pincus, Widiger, Davis, & First, 1990; Kendler, 1990). For instance, the ancient Greeks and Romans developed a rather crude psychiatric diagnostic system based around five different categories: "phrenitis, mania, melancholia, hysteria, and epilepsy" (Frances et al., 1990, p. 1440). The science of systematically categorizing disease states is nosology, and thus the science of categorizing mental and behavioral abnormalities is called psychiatric nosology (Kendler, 1990).

Kendler (1990) observed that, historically, the creation of a psychiatric nosology in post-Renaissance Western culture up to the mid-20th century was dominated by two different techniques: initially, there was "the great professor principle" (p. 969), which eventually gave way to "the consensus of experts" (p. 969). The method of creating a psychiatric nosology through "the great professor principle" involved a single clinician and/or researcher who was highly respected in the psychiatric field creating a nosological system based on his own clinical observations and synthesis of the available literature. Unfortunately, different "great professors" came up with different diagnostic criteria for similar disorders and there was little consensus even within the same country as to how to define various psychiatric disorders (Kendler, 1990).

"The consensus of experts" method was thought to be an improvement on the "the great professor principle" in that, instead of a single individual, a committee of experts created a nosological system by way of consensus and/or majority vote (Kendler, 1990). In the 20th century, there arose a number of competing psychiatric diagnostic systems in North America and Europe that were created by the "consensus of experts." Most notable among these were the Feighner Diagnostic Critiera (FDC; Feighner et al., 1972), the Research Diagnostic Criteria (RDC; Spizter, Endicott, & Robins, 1978), the section on psychiatric disorders in the World Health Organization's (WHO) *International Classification of Disease (6th Revision)* (ICD-6; WHO, 1948, as cited by APA, 1987), ICD-7 (WHO, 1955, as cited by APA, 1987), ICD-8 (WHO, 1969, as cited by APA, 1987), ICD-9 (WHO, 1977, as cited by APA, 1987), the *Diagnostic and Statistical Manual of Mental Disorders* (1st Edition) (DSM-I; APA, 1952, as cited by APA, 1987) and DSM-II (APA, 1968, as cited by APA, 1987). However, there were important differences between these different systems, and none of them reached anything resembling acceptance as a universal standard for psychiatric diagnoses (Kendler, 1990; Philipp, Maier, & Delmo, 1991a, 1991b, 1991c). Thus, it became evident that there was a need for a universally agreed upon standard for psychiatric nosology.

DSM-III and beyond

In North America, a consensus on an "official" psychiatric diagnostic system was not finally achieved until 1980 with the publication of the *Diagnostic and Statistical Manual of Mental Disorders* (3rd Edition) (DSM-III) by the American Psychiatric Association (Barlow & Durand, 2005; Kendler, 1990; Frances et al., 1990). Unlike previous psychiatric nosological systems, DSM-III managed to achieve universal

acceptance in the health professions as a standard psychiatric diagnostic system (Barlow & Durand, 2005; Kendler, 1990; Frances et al., 1990).

The format of DSM-III consisted of a catalogue of a large number of psychiatric disorders, each of which was defined by a set of criteria that were labeled "A," "B," "C," etc. (APA, 1980). The individual diagnostic criteria for a disorder consisted of either a symptom list or one or more inclusion or exclusion criteria (APA, 1980). The inclusion and exclusion criteria specified additional conditions beside a set of core symptoms that defined who could or could not be diagnosed with a disorder. The diagnostic criteria themselves were identified by a mixture of historical tradition, clinical observation, and systematic research (Clark, Watson, & Reynolds, 1995; Frances et al., 1990; Kendler, 1990). DSM-III used a dichotomous categorical approach toward diagnosis, that is, an individual either had a particular disorder or did not (APA, 1980; Clark et al., 1995). Disorders that were hypothesized to be similar in nature were grouped into families, e.g., the substance related disorders, the schizophrenias, mood disorders, anxiety disorders, etc. (APA, 1980). The symptom criteria for each disorder was either monothetic, i.e., all symptoms out of a total set needed to be fulfilled in order to meet the requirements for a formal diagnosis of a particular disorder, or polythetic, i.e., only some of the symptoms out of a total set needed to be fulfilled in order to meet the diagnostic requirements for a particular disorder (Clark et al., 1995; Frances et al., 1990; Widiger & Trull, 1991). In DSM-III, the monothetic approach was more heavily favored in defining disorders than the polythetic approach (Clark et al., 1995; Frances et al., 1990; Widiger & Trull, 1991).

The final codifications of the psychiatric diagnostic criteria in DSM-III were achieved by way of committee (APA, 1980, 1987; Kendler, 1990). However, Kendler (1990) notes that

> the efforts of this committee of experts differed from those of their predecessors in two important ways. First, the committee made a conscious effort to use available "scientific" information in the development and evaluations of proposed nosologic changes. Second, they decided to require explicit diagnostic criteria that would greatly facilitate future studies of their reliability and validity (pp. 969-970).

In other words, according to Kendler, rather than being a nosological system that was created by "great professors" or "a committee of experts," DSM-III was a true "scientific nosology" (p. 970) in that the scientific method was used in creating the final diagnostic criteria. As proof of the successful implementation of "scientific nosology" in creating DSM-III, Kendler notes that DSM-III did have an overall average increase in the reliability and validity of various psychiatric diagnoses over previous nosological systems, though this was seen more in some types of disorders than for others.

DSM-III was also noteworthy in that it was atheoretical in its orientation, that is, there was no influence of a particular school of thought or scientific paradigm of mental disorders, e.g., psychoanalytic, cognitive, behavioral, humanistic, etc (Clark et al., 1995; Frances et al., 1990; Widiger & Trull, 1991). This lead to a more or less descriptive approach in defining different disorders, that is, disorders were defined without any reference to their etiology or prognosis (Frances et al., 1990).

Although DSM-III gained widespread acceptance, even beyond North America, it was subject to much criticism about its reliability and validity (Barlow & Durand, 2005; Clark et al., 1995). In response to these criticisms, a revised edition was issued, DSM-IIIR (APA, 1987). One of the major changes in DSM-IIIR was that many monothetic criteria for various disorders were changed to polythetic criteria (Clark et al., 1995). DSM-IIIR was eventually replaced by DSM-IV (APA, 1994), which was considered a major overhaul of DSM-IIIR (Frances et al., 1990). DSM-IV was subject to a text revision, which was published as the DSM-IV Text Revision[1] (DSM-IV TR; APA, 2000). DSM-IV TR corrected, updated and revised the text of DSM-IV in light of the psychopathological literature that was published since the publication of DSM-IV, however, "no substantive changes in the criteria sets were considered, nor were any proposals entertained for new disorders [or] new subtypes …" (APA, 2000, p. xxix).

The threshold for altering diagnostic criteria from DSM-IIIR to DSM-IV was set quite high in that a greater amount of evidence from the scientific literature and/or field trials was needed to revise a set of diagnostic criteria than was the case for previous editions of DSM (Frances et al., 1990; Kendler, 1990). Thus, DSM-IV is the most current set of official psychiatric diagnostic criteria in North America, and for better or worse, modern psychiatric nosology in North America is officially represented by it. It should be noted that the first set of committees have been put together by the APA to begin the process of revising DSM-IV into DSM-V, which is expected to be released in approximately 2011 (Barlow & Durand, 2005).

[1] For ease of reference, in the current review both DSM-IV and DSM-IV TR will be generically referenced as DSM-IV.

The DSM-IV nosological system is widely accepted as the de facto "gold standard" for psychiatric diagnosis in North America, however, much like its predecessors, it has been subject to criticism. Nathan and Langenbucher (1999) have reviewed some of the major criticisms of DSM-IV, the most prominent of which is that not all the disorders have high reliability. There is also a concern that, at least for some of the disorders, validity of the diagnostic criteria may have been sacrificed in favor of easier clinical utility (Nathan & Langenbucher, 1999).

Clark et al. (1995) present evidence that the categorical nature of DSM may be intrinsically flawed and inadequate. One problem for the categorical nature of DSM is comorbidity, that is, certain disorders tend to cluster more frequently in patients, e.g., it has been repeatedly observed that anxiety and depression often co-occur in patients (Krueger & Piasecki, 2002). Another problem identified by Clark et al. is that with an increase in the use of polythetic criteria in DSM-IIIR and DSM-IV for most disorders, there is a concomitant increase in the heterogeneity of symptoms for individuals diagnosed with the same disorders. Finally, Clark et al. note that many researchers have identified potential problems with DSM's "phenomenological organization" (p. 139) in that the assignment of certain disorders into particular families of disorders is incorrect because of logical, theoretical and/or empirical reasons.

The DSM diagnostic criteria were ultimately born out of a set of compromises that balanced the needs of researchers versus the needs of clinicians, ease of use versus comprehensiveness of diagnostic criteria, the needs of mental health professionals such as psychiatrists, clinical psychologists, school psychologists, counselors, and social workers versus nonmental health professionals such as public policy makers, law enforcement,

agents of insurance companies (Clark et al., 1995; Frances et al., 1990; Nathan & Langenbucher, 1999). Despite the problems with the current DSM system, for the foreseeable future, it will continue to be the "gold standard" for mental and behavioral disorders in North America and even beyond (Barlow & Durand, 2005).

Depression

One area of psychiatric nosology that is especially prominent is depression, which is otherwise known as major depressive disorder (MDD; APA, 2000). Depression is a highly pervasive disorder in our society. Wittchen, Knauper and Kessler (1994) reviewed major studies that calculated prevalence rates of depression and found a median value of 16.1% lifetime rates of depression using data from all studies. One of the most serious mental health consequences of depression is that it leads to an increase in the probability to commit suicide (Barlow & Durand, 2005). It also has serious psychobiological consequences, such as having a direct impact on brain areas associated with memory and related cognitive functions (Shors & Leuner, 2003). The diagnostic criteria for depression should be as reliable and valid as possible since they can help identify individuals who need treatment as well as lead to improved psychometric measures for depression, which would be enormously useful in advancing the study of depression.

As a universally acknowledged disorder, depression has had a central place in all modern psychiatric nosological systems (Parker, 2005; Philipp et al., 1991a, 1991b, 1991c). The immediate historical predecessor to the DSM criteria for MDD was the Feighner criteria (Feighner et al., 1972), also known as the Washington University criteria (Spitzer et al., 1978; Zimmerman, McGlinchey, Young, & Chelminski, 2006b). The set of depression symptoms in the Feighner criteria was in turn based on a set of

depression criteria developed by Cassidy et al. (1957, as cited by Zimmerman, McGlinchey, Young, & Chelminski, 2006a).

The Feighner criteria (Feighner et al., 1972) stated that for a diagnosis of depression the following must be observed:

> A. Dysphoric mood characterized by symptoms such as the following: depressed, sad, blue, despondent, hopeless, "down in the dumps," irritable, fearful, worried, or discouraged.
>
> B. At least five of the following criteria are required for "definite" depression; four are required for "probable" depression. (1) Poor appetite or weight loss … (2) Sleep difficulty (include insomnia or hypersomnia). (3) Loss of energy … (4) Agitation or retardation. (5) Loss of interest in usual activities or decrease in sexual drive. (6) Feelings of self-reproach or guilt (either may be delusional). (7) Complaints of or actually diminished ability to think or concentrate … (8) Recurrent thoughts of death or suicide, including thoughts of wishing to be dead. (p. 58)

In addition to criteria A and B, there were also temporal and exclusionary criteria, which state that an individual had to have noticeable levels of distress for a month and no severe preexisting psychiatric conditions of a nondepressive nature.

Since the publication of the Feighner criteria, the criteria for MDD have gone through several different iterations, including the different versions of DSM. The most current criteria for MDD are found in DSM-IV (APA, 2000). DSM-IV does not explicitly define symptom criteria for MDD, but instead, it defines symptom criteria for a major depressive episode (MDE), which is needed to fulfill the diagnostic criteria for

MDD. In essence, the MDE is considered a "building block" (APA, 2000, p. 345) for a diagnosis of MDD. The key symptom criteria for a MDE are:

> Five (or more) of the following symptoms have been present during the same 2-week period and … at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure….
>
> 1. depressed mood most of the day, nearly every day …
> 2. markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day …
> 3. significant weight loss … or weight gain … or decrease or increase in appetite nearly every day …
> 4. insomnia or hypersomnia nearly every day
> 5. psychomotor agitation or retardation nearly every day …
> 6. fatigue or loss of energy nearly every day
> 7. feelings of worthlessness or excessive or inappropriate guilt … nearly every day …
> 8. diminished ability to think or concentrate, or indecisiveness, nearly every day …
> 9. recurrent thoughts of death … , recurrent suicidal ideation … , or a suicide attempt or specific plan for committing suicide (APA, 2000, p. 356)

DSM-IV requires that the above symptoms must cause "clinically significant distress" (APA, 2000, p. 356). DSM-IV also has additional exclusionary criteria for a MDE, which are that the above symptoms must not be due to drugs or bereavement or other nonpsychiatric medical problems (APA, 2000). A diagnosis of MDD requires the

presence of one or more MDE(s) in the recent past that were not concurrent with psychotic disorders, such as schizophrenia, and the absence of a manic episode (APA, 2000). Furthermore, the diagnosis of MDD has several specifiers to enhance the description of MDD, most noteworthy of which is the "melancholic features specifier," (APA, 2000, p. 419) which is defined as a "loss of interest or pleasure in all, or almost all, activities or a lack of reactivity to usually pleasurable stimuli," (APA, 2000, p. 419) and the "atypical features specifier," (APA, 2000, p. 420) which has "historical significance (i.e., atypical in contradistinction to the more classical 'endogenous' presentations of depression)" (APA, 2000, p. 420).

Not surprisingly, much of the controversies concerning the DSM criteria for depression mirror the controversies for DSM in general, i.e, a lack of solid theoretical underpinnings for the diagnostic criteria, heterogeneity of symptoms in individuals diagnosed with MDD, and issues with the reliability and validity of MDD (Carroll, 1984). There have also been criticisms specific to the construct of depression, the most prominent of which is that depression is not a homogenous construct (Parker, 2000, 2005; van Praag, 2001). Certain clinicians and researchers believe that depression should not be conceived as a single monolithic construct, and instead, the construct of depression should be divided up into subtypes, which presumably would increase the reliability and validity of the diagnostic construct of depression (Parker, 2000, 2005; van Praag, 2001). However, among researchers and clinicians who favor such an approach, there is much controversy as to how to divide up depression into subtypes (Parker, 2000, 2005; van Praag, 2001).

Psychometric Analyses of the Construct of Depression

Zimmerman et al. (2006a, 2006b) observed that one of the key issues with the diagnostic criteria for depression is that little effort has gone into systematically evaluating and refining the criteria using psychometric techniques during their initial development and subsequent release. In fact, it is somewhat remarkable that psychodiagnostic criteria in general, at least as represented by the nosological systems of FDC, RDC, ICD and DSM, have remained relatively untouched by sophisticated psychometric modeling (Zimmerman et al., 2006a, 2006b). While there have been a few notable recent exceptions to this trend (for instance, Aggen, Neale, & Kendler, 2005; Langebucher et al., 2004; Reiser, 1989), most studies that have evaluated various psychodiagnostic criteria have not used psychometric approaches (Zimmerman et al., 2006a, 2006b). Commenting on their own effort to psychometrically model and evaluate the DSM-IV diagnostic symptom criteria for depression, Zimmerman et al. (2006b) stated that their study

> is about 30 years too late. Ours is the type of methodical psychometric analysis that should have been conducted when initially developing the sets of diagnostic criteria. Nonetheless, it remains relevant to determine whether the assumptions underlying the diagnostic rules have empirical support, and to examine the impact of these rules on clinical practice. (p. 153)

Considering the weight attached in our society to psychiatric nosological diagnostic criteria in defining what is considered "normal" and "abnormal" behavior, the lack of systematic psychometric analysis of these criteria appears to be an important and glaring oversight. The psychometric modeling of psychodiagnostic criteria could provide an

important additional perspective on many of the controversies in psychiatric nosology. The use of formal psychometric models could help resolve some of the ongoing debates about the nature of psychodiagnostic criteria, or at the very least, provide the framers of future editions of DSM more sophisticated tools that can be used to guide the revision of diagnostic criteria. Given the importance of depression in any psychiatric nosological system, its symptom criteria deserve a high degree of scrutiny with regard to their reliability and validity.

Purpose of Dissertation

The purpose of the present study is to psychometrically assess the DSM-IV symptom criteria for a MDE. Psychometric theory can be divided into two main branches: classical test theory and item response theory, which is also considered to be modern test theory (Crocker & Algina, 1986). There have been a small number of studies that have investigated the diagnostic symptom criteria for depression using either a classical test theory paradigm (Buchwald & Rudick-Davis, 1993; Faravelli, Servi, Arends, & Strik, 1996; Zimmerman et al., 2006a, 2006b) or an item response theory paradigm (Aggen et al., 2004; Reiser, 1989). However, most of these studies have some sort of important limitations, including small datasets (Buchwald & Rudick-Davis, 1993; Faravelli et al., 1996), using datasets that are unrepresentative of the population in Western culture (Aggen, Neale, & Kendler, 2005), and inadequate implementation of classical test theory (Buchwald & Rudick-Davis, 1993; Faravelli et al., 1996; Zimmerman et al., 2006a).

Despite their limitations, these studies have discovered some potentially serious problems with the diagnostic symptom criteria for depression. Serious questions remain

about: (1) how reliable are the individual symptom criteria for depression, (2) whether individual symptom criteria measure the same diagnostic construct of depression, (3) how much information individual symptom criteria contribute to a diagnosis of depression, and (4) how effective are individual symptom criteria in differentiating individuals who have a low, moderate, or severe instance of depression.

The proposed project will attempt to overcome some of the limitations of previous psychometric studies on the diagnostic symptom criteria of depression. First, all analyses will be done on a nationally representative community sample of individuals from the United States by using data from the National Comorbidity Study-Replication, which was a large scale psychiatric epidemiological study (Kessler & Merikangas, 2004; Kessler et al., 2004). Second, a comprehensive set of psychometric analyses will be carried out on the nine core symptoms of a MDE found in DSM-IV (APA, 2000). The psychometric analyses will consist of techniques drawn from both classical test theory and item response theory. The psychometric analyses will provide an insight into the contribution of each individual diagnostic symptom towards a diagnosis of depression by addressing the four major questions listed above.

For purposes of the psychometric analyses, the nine symptoms of depression will be treated as individual items. As reviewed above, an individual needs five out of the nine key symptoms in order to meet the requirements for a diagnosis of MDD in the DSM-IV system. Even though the diagnosis of MDD is categorical, the underlying symptom criteria are polythetic, which implies that they can be conceptualized as a set of items on a standardized psychological test that measure an individual's place on the continuum of a latent construct. The final count of symptoms used to determine whether

an individual meets the criteria for a diagnosis of MDD is equivalent to a summed total test score of dichotomous items (Reiser, 1989). The requirement that five out of the nine symptoms must be present for a diagnosis of MDD is equivalent to a cutpoint on a continuous psychometric scale. These features of the symptoms in the DSM-IV diagnostic criteria for MDD make them amenable to a formal psychometric analysis.

The DSM criteria were not initially created from a psychometric perspective, so this study is unavoidably somewhat awkward in its research design because it essentially is "reverse engineering" a set of latent construct indicators that were not initially created with psychometric validation in mind. The verbal content of the diagnostic criteria is somewhat inelegant from a psychometric perspective (Anastasi & Urbina, 1997; Crocker & Algina, 1986; Cronbach, 1984) because many of the criteria have a compound structure that simultaneously examines opposite behavioral tendencies in the same domain, e.g., loss or gain of weight. Given such awkwardly constructed indicator items and the associated diagnostic algorithms, the setup for a psychometric analysis is less than optimal. However, for better or worse, the DSM-IV diagnostic criteria for the MDE are used to construct the "official" operational definition of depression for the North American mental health professions, and therefore the diagnostic criteria of depression deserve a close scrutiny of how they are actually performing in the general population.

The second chapter of this dissertation proposal will review the relevant literature on psychometric theory and the application of psychometric theory to the diagnostic criteria of depression. The first section of the second chapter will review the fundamental theoretical basis and history of measurement theory. Following the theoretical and historical review of measurement theory, there will be a review of classical test theory,

item response theory, and their respective advantages and disadvantages with regard to their application to noncognitive constructs such as those found in the areas of psychopathology, personality and attitudes. The chapter will then explore several key issues that have been identified as serious concerns in the application of item response theory to the noncognitive measures. The last section of the second chapter will review the application of the classical test theory and item response theory paradigms to the diagnostic symptom criteria for depression. Methodologically, the paper will focus more on psychometric models for dichotomous items, as opposed to polytomous items, partially because the DSM diagnostic symptom criteria are dichotomous and partially also for the sake of brevity.

The third chapter will focus on the methodology of the proposed study. It will consist of two main sections. The first section will be a description of the National Comorbidity Study – Replication (Kessler & Merikangas, 2004; Kessler et al., 2004), which is the source of the data set for the proposed study, and the Composite International Diagnostic Interview (Kessler & Üstün, 2004), which was the main instrument that was utilized in the National Comorbidity Study – Replication. Included in the review of the National Comorbidity Study – Replication will be a description of some key methodological issues concerning the National Comorbidity Study – Replication that impinge on the proposed study. The second section of the third chapter will be a description of the set of psychometric analytic techniques that will utilized in the proposed study.

The fourth chapter will contain the results of psychometric analyses of the nine symptom criteria of a MDE. The first part will present the results of the factor analysis,

which will determine the underlying dimensional structure of the symptom criteria. The second part will consist of the results of the classical test theory analysis of the symptom criteria, which will include reliability values, item difficulty values, and item-total score correlations. The third part will consist of the results of the item response theory analysis, both in table and graphical form.

The fifth chapter will consist of a discussion of the results of the psychometric analyses. The discussion will go over some of the clinical implications of the psychometric analyses, including an examination of the more salient and consistent patterns of symptoms in individuals with depression, an evaluation of the five symptom cutpoint required for a diagnosis of depression, and an overview of how the symptoms of depression can evolve in a patient. The chapter will also contain a discussion of the future avenues of research that are suggested by the results of the psychometric analyses.

CHAPTER II. LITERATURE REVIEW

Introduction

Measurement Theory

The measurement of the properties of objects or observed phenomena is a fundamental issue in the physical, biological and behavioral sciences (Crocker & Algina, 1986; McDonald, 1999). Measurement formally involves the "assignment of numbers to an attribute [or property] according to a rule of correspondence" (McDonald, 1999, p. 55). The assignment of numbers to an attribute or phenomena allows for it to be quantified and hence ordered on a meaningfully defined scale or continuum (Crocker & Algina, 1986; McDonald, 1999).

Many of the attributes of phenomena that are of interest in the behavioral sciences, such as intelligence or personality, are not as directly observable, and hence not easily measurable, as they are in the physical and biological sciences. Psychological attributes are often considered to be latent constructs, which are "hypothetical concepts – products of the informed scientific imagination of social scientists who attempt to develop theories for explaining human behavior" (Crocker & Algina, 1986, p. 4). To measure such unobservable psychological attributes, behavioral scientists first need to operationally define the latent constructs, which involves identifying observable, and hence measurable, behaviors that are considered to be indicators of the latent constructs (Christensen, 1988). For example, the property of intelligence can not be directly assessed, but cognitive and developmental psychologists have identified tasks on which successful performance should theoretically be influenced by intelligence and therefore these tasks can be used to indirectly assess the level of intelligence in an individual. In

most instances, psychological latent constructs are assessed through psychological tests, which consist of a set of tasks that are generically known as items. A psychological test can also be referred to as a scale and in this paper a test or scale will be defined as a set of items that are usually associated with a single latent construct. Note that a published test can have multiple scales that are essentially subtests.

Psychometrics: A Brief Historical Overview of Classical Test Theory and Item Response Theory

The science of psychological tests is known as psychometrics. Psychometrics had its origins in the mid to late 1800's as part of an effort to quantify cognitive and affective attributes in humans (Anastasi & Urbina, 1997; Crocker & Algina, 1986). By the late 1920's, researchers in psychology and education systematically developed and formalized the field of psychometrics into a paradigm that today is known as Classical Test Theory (CTT; Anastasi & Urbina, 1997; Baker, 2001; Crocker & Algina, 1986).

CTT provides a comprehensive framework for creating, evaluating and scoring scales and their associated items that is, for the most part, based on introductory level statistical concepts such as moments, correlation, regression and ordinary least squares. The set of items that make up a scale are evaluated and refined by CTT procedures so that they are usually measuring one specific latent construct. Scoring a test in the CTT framework typically involves summing the scores of individual items that are part of the scale(s) that make up the test. One of the advantages of CTT is that its techniques are mathematically and computationally tractable without the benefit of computers.

The initial theoretical groundwork for an alternative to the CTT paradigm known as Item Response Theory (IRT) was laid down in the 1940's by D. N. Lawley (Baker,

2001; McDonald, 1999). A more comprehensive theoretical framework for IRT was subsequently developed and shaped from the 1950's to the 1970's by a number of prominent psychometricians and statisticians, including Frederick Lord, Georg Rasch and Benjamin Wright (Baker, 2001; McDonald, 1999). The mathematics and computational algorithms of IRT are far less tractable than CTT and therefore the practical implementation of IRT had to wait until the 1970's and 1980's for the development of computers that could carry out the necessary computations for IRT.

The key feature of IRT that sets it apart from CTT is the explicit mathematical modeling of the stochastic relationship between the performance on an individual item and the underlying continuous scale of the latent construct that the item is theorized to be measuring (Baker, 2001; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). IRT is considered to be a major advancement over CTT because it allows for a more precise analysis of individual items and how they relate to the underlying latent construct. As a result, the use of IRT can lead to better tests that more accurately measure a construct across specific targeted areas of the latent scale continuum than CTT (Embretson & Hershberger, 1999; Embretson & Reise, 2000; Hambleton et al., 1991); but see Fan (1998) for a contrarian view.

Applications of CTT versus IRT

Historically, CTT has been applied both to cognitive tests, such as academic aptitude and intelligence measures, and noncognitive tests, such as personality, clinical and attitudinal measures (Crocker & Algina, 1986; Cronbach, 1984). IRT, however, has been mostly applied to cognitive tests and only sporadically applied to noncognitive tests (Embretson & Hershberger, 1999; Embretson & Reise, 2000; Reise, 1999). This

historical trend may be due, in part, to IRT initially being developed by psychometricians and researchers working in the educational and aptitude testing fields (Embretson & Reise, 2000). Indeed, much of the technical jargon associated with IRT still is much more appropriate for cognitive testing, e.g., terms such as "item difficulty" (Rouse, Finger, & Butcher, 1999). However, recent reviews by Embretson and Reise (2000) and Reise and Henson (2003) demonstrate that although CTT techniques still dominate the field of noncognitive assessment, there is a growing interest among psychologists in applying IRT to noncognitive tests since the early 1990's.

Noncognitive measures are an important concern in the behavioral sciences since they encompass a wide variety of psychological tests that have a pervasive use in our society (Cronbach, 1984). Noncognitive measures are extensively used in all age ranges for mental health screening and for tracking the success of mental health treatments. They are also used for job screening and measuring attitudes, and as tools for the creation, implementation, and evaluation of certain public policy initiatives. The creation of accurate measurement instruments for noncognitive attributes is therefore essential.

Classical Test Theory

CTT Models

As noted above, the core framework of CTT was formulated by the 1920's and since then a vast literature has been spawned containing numerous developments and refinements concerning the theory and application of CTT to many areas of psychology and education (Crocker & Algina, 1986). The current review of CTT will necessarily be terse and highlight those points which are most essential in differentiating CTT from IRT with regard to their application to noncognitive measures.

CTT is based on true score theory, which assumes that the summed test score of a scale is a random variable that is the sum of two parts, the true score and error,

$$X = T + e, \tag{1}$$

where X is the observed test score, T is the true score and e is the error (Crocker & Algina, 1986). T is "defined as the expected (average) score an individual would receive if they were repeatedly administered parallel measures an infinite number of times. Simply stated, two measures are considered parallel if the true score variance is equal across both measures" (Reise & Henson, 2003, p. 93). The error is the difference between T and an observed test score and it is assumed to be uncorrelated with the true score (Crocker & Algina, 1986; DeVellis, 2006; Kline, 1998).

Psychometricians have developed standards for what constitutes a "good" CTT-based psychological test. Foremost among these standards is the property of high reliability (Kline, 1998). The reliability for a set of scores is theoretically derived by Crocker and Algina (1986) as:

$$\rho_{X_1X_2} = \frac{\sigma_T^2}{\sigma_X^2}, \tag{2}$$

where $\rho_{X_1X_2}$ is a "reliability coefficient" (p. 116), σ_T^2 is the variance of the true score, and σ_X^2 is the variance of the observed score. The formula for $\rho_{X_1X_2}$ shows that reliability is the proportion of observed score variance that can be explained by true score variance.

Because the true score is unobserved, a variety of methods have been developed to assess reliability based on the observed scores of a group of examinees, either from one or more test taking sessions and/or forms. One well known method that uses two sets of test scores is test-retest reliability, which involves giving the test at two different occasions, or an alternate form of the test on the same occasion, and computing the correlation coefficient for scores from either both sessions or both forms (Crocker & Algina, 1986). A test-retest correlation of .7 is usually considered the minimum acceptable value for good test-retest reliability (Crocker & Algina, 1986; Kline, 1998). Note that the mathematical symbol for the reliability coefficient ($\rho_{X_1X_2}$) implies that it is a correlation between two different sets (x_1 and x_2) of test scores.

One of the most popular methods for assessing reliability of test scores from a single test taking session or form is Cronbach's α (Crocker & Algina, 1986). Several other techniques of assessing reliability of test scores from a single test taking session are essentially special cases of Cronbach's α, which most likely contributes to its popularity (Crocker & Algina, 1986).

Cronbach's α is a measure of the consistency of item scores with each other (Crocker & Algina, 1986). Crocker and Algina (1986) derive the computation formula of Cronbach's α as:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right), \tag{3}$$

where $\hat{\alpha}$ is Cronbach's α, k is the total number of items for a particular test, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the variance of the total summed test score. The value of

Cronbach's α ranges from 0 to 1. Cronbach's α will be high in situations where items in a test all have high intercorrelations with each other (DeVellis, 2006; Kline, 1998). In other words, a high Cronbach's α for a set of test scores indicates that the scores of items from that test show a similar pattern of responding for each individual test taker relative to other test takers in the group.

The reliability of test can be converted to a standard error of measurement (SEM) for a given set of test scores. The formula for the SEM is:

$$SEM = \sigma_X \sqrt{1-\rho}\,, \tag{4}$$

where σ_X is the standard deviation for a given set of test scores, and ρ is an estimate of the reliability of the test (Anastasi & Urbina, 1997; Crocker & Algina, 1986).

The above standards for a good test dealt with the total test score, but there are also item level standards for a good test in the CTT paradigm. One item level statistic that is important in determining test quality is item difficulty, p_i, which for the *i*th dichotomous cognitive item is the proportion of individuals that respond correctly, or for *i*th dichotomous noncognitive item is the proportion of individuals that respond positively (Anastasi & Urbina, 1997; Crocker & Algina, 1986). Under CTT, a psychological test is considered optimal in terms of maximizing the total true score variance if all the test's p_i values fall in a range between .30 to .70, that is, if the items are considered moderately difficult (Anastasi & Urbina, 1997; Crocker & Algina, 1986). Another item level statistic is the item-total score correlation (ρ_{XT}), which should be high for a good item (Anastasi & Urbina, 1997; Crocker & Algina, 1986). ρ_{XT} is also known as the item discrimination index, because it is a measure of how much an item can discriminate between the

individuals with a low and high total test score (Anastasi & Urbina, 1997; Crocker & Algina, 1986).

Another important standard for a good psychological test under CTT is that each individual scale possesses unidimensionality, that is, the item scores from a scale are indicative of only one latent construct (DeVellis, 2006). Unidimensionality in CTT is tested with factor analysis, either of the exploratory and/or confirmatory variety (see Gorsuch (1983; 1997), Hayduk (1987), Stevens (1996), and Tabachnick and Fidell (2006) for a comprehensive review of exploratory and confirmatory factor analysis). Unidimensionality is an ideal to aim for; however, most tests cannot be perfectly unidimensional (Reckase, 1979).

Advantages and Disadvantages of CTT Models for Noncognitive Measures

Advantages of CTT

Because CTT was developed early in the history of psychometrics, it has been extensively applied in the construction of both cognitive and noncognitive tests. DeVellis (2006) notes that CTT has some important advantages for researchers who develop and use noncognitive measures: (1) CTT uses statistical concepts with which most researchers are familiar, i.e., the kind of statistics taught in introductory level application oriented statistics courses. (2) CTT analyses can usually be done with most commercially available statistics software packages. (3) The true score model (Equation 1) of CTT appears to adequately work for many constructs in the behavioral and health science fields. (4) The relationships between item scores and the total test score do not have to be "optimal" (p. 57), i.e., ρ_{XT} values for a set of test items can be moderate,

however, in such cases a large number of items is needed to adequately capture a construct.

Disadvantages of CTT

In contrast to its advantages, DeVellis (2006), Hambleton et al. (1991), Reise (1999), Waller, Tellegen, McDonald and Lykken (1996), Weiss (1995), and Wilson, Allen and Li (2006a, 2006b) note that CTT does have serious limitations, among the most prominent of which are: (1) CTT parameters are not invariant, that is, they are dependent on the characteristics of the sample on which test data were collected. (2) Because reliability in CTT is directly correlated with the number of items in a test, tests constructed in the CTT paradigm often have a large number of items that can often be redundant in their content. (3) CTT has only one measure of reliability for a given set of scores of a test. (4) The CTT paradigm encourages the construction of tests with items that often have on average a moderate level of difficulty and therefore they are best in discriminating individuals who are in the middle range of the underlying latent construct scale. (5) Since scoring in CTT usually involves a simple summation of the item responses, the contribution of each item to the total score is not weighted by an item's individual relationship to the construct. (6) In contrast to the optimistic position of DeVellis that was described above about the utility of the CTT model in noncognitive test development, Waller et al. argue that the CTT model is inappropriate for many noncognitive scales in psychology where there is a nonlinear relationship between the latent scale and the test items because the CTT model is an inherently linear statistical model. To be fair to DeVellis's position, though, it should be noted that he comes from the perspective of health research, and according to DeVellis, in the area of health

research there are indeed many scales with approximately continuous items that are more likely to have a linear relationship with a latent construct.

Item Response Theory

IRT Models

As with the review of CTT above, the current review of IRT will necessarily be terse and will highlight those characteristics of IRT which make it a more powerful tool for psychometric modeling as compared to CTT. The key concept in the framework of the IRT psychometric paradigm is that the relationship between the probability to respond to an item and the latent scale (θ) of a psychological construct is mathematically represented by a nonlinear equation, typically logistic, which is known as an item characteristic curve (ICC) or as an item response function (IRF) (Rouse et al., 1999). In IRT, each *i*th item of a scale is assigned an ICC, which can than be plotted on a graph with θ on the x-axis and probability to respond to the item on the y-axis (Embretson & Reise, 2000; Hambleton et al., 1991). For dichotomous items, there are three standard forms of the ICC that are the most well-known: the one-parameter (1PL) model, the two-parameter (2PL) model and the three-parameter (3PL) model (Embretson & Reise, 2000; Hambleton et al., 1991).

The equation of the 1PL model is:

$$P(X_{is} = 1 \mid \theta_s, \beta_i) = P_i(\theta) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)}, \quad (5)$$

where $P_i(\theta)$ is the probability to respond to item *i*, θ_s is the location of subject *s* on the latent scale, and β_i is the point on the latent scale where subjects have a .5 probability of responding to item *i* (Embretson & Reise, 2000; Hambleton et al., 1991). β_i is otherwise

known as the difficulty parameter and can also be referred to as the *b* parameter (Embretson & Reise, 2000; Hambleton et al., 1991). A test whose items can be graphed by the 1PL model, i.e., the item ICCs differ only in their level of difficulty, is known as a Rasch model or scale, which implies that the total score is a sufficient statistic to use in reporting performance on a test (Embretson & Reise, 2000; Hambleton et al., 1991).

The equation of the 2PL model is:

$$P(X_{is}=1 \mid \theta_s, \beta_i, \alpha_i) = P_i(\theta) = \frac{\exp(\alpha_i(\theta_s - \beta_i))}{1+\exp(\alpha_i(\theta_s - \beta_i))}, \quad (6)$$

where $P_i(\theta)$ and β_i are defined as above for the 1PL model, and α_i is the slope of the ICC at the point on the latent scale corresponding to the β_i parameter (Embretson & Reise, 2000; Hambleton et al., 1991). α_i is known as the discrimination parameter because it models how effective the ICC is in separating test takers who have θ values on the latent scale below and above the point corresponding to the β_i parameter (Embretson & Reise, 2000; Hambleton et al., 1991). α_i can also be referred to as the *a* parameter.

The equation of the 3PL model is:

$$P(X_{is}=1 \mid \theta_s, \beta_i, \alpha_i, \gamma_i) = P_i(\theta) = \gamma_i + (1-\gamma_i)\frac{\exp(\alpha_i(\theta_s - \beta_i))}{1+\exp(\alpha_i(\theta_s - \beta_i))}, \quad (7)$$

where $P_i(\theta)$, β_i, and α_i are defined as above for the 2PL model, and γ_i is an intercept term known as the guessing parameter (Embretson & Reise, 2000; Hambleton et al., 1991). The term guessing parameter has its origin in the modeling of cognitive tests and it represents the probability that an individual with a low θ level can get an item right simply by guessing (Rouse et al., 1999). γ_i can also be referred to as the *c* parameter.

The computation of the ICC parameters for a set of items in a scale is known as calibration and is usually accomplished by using maximum likelihood estimation techniques (see Embretson & Reise, 2000, for more detail). Once the ICC parameters are known, maximum likelihood techniques can then be used to calculate a score, known as θ, for each examinee based on their responses to the set of items in a scale (Embretson & Reise, 2000; Hambleton et al., 1991).

There are three key assumptions of IRT: monotonicity, unidimensionality, and local independence (McDonald, 1999; Sijtsma & Molennar, 2002). *Monotonicity* means that the relationship between the latent construct and an indicator is positive, i.e., $P_i(\theta)$ is continuously increasing as θ is increasing (McDonald, 1999; Sijtsma & Molennar, 2002). The concept of *unidimensionality* for a set of items in IRT is essentially the same as the concept as unidimensionality in CTT (McDonald, 1999). The assumption of *local independence* is met when responses to one item in a set of items are not related to responses to other items in the set, once the influence of the latent construct is taken into account (McDonald, 1999). There is a strong and weak version of local independence (see McDonald, 1999, for more detail). Weak local independence is adequate for CTT, but strong local independence is necessary for IRT (McDonald, 1999). The assumptions of IRT models are essentially identical irrespective of whether an IRT model is applied to a cognitive or noncognitive measure (Embretson & Reise, 2000; McDonald, 1999; Sijtsma & Molennar, 2002). A comprehensive review of the issues concerning the assumptions of IRT is beyond the scope of the paper, but thoughtful reviews may be found in Hattie (1985), Gorsuch (1997), and Embretson and Reise (2000)

An important feature of IRT is that the ICC can be transformed into an item information function (IIC). The formula for the IIC is:

$$I(\theta) = \frac{P_i^*(\theta)^2}{P_i(\theta)(1 - P_i(\theta))}, \tag{8}$$

where $P_i(\theta)$ is the ICC for item *i* and $P_i^*(\theta)^2$ is the first derivative squared of the ICC for item *i* (Embretson & Reise, 2000; Reise & Henson, 2003; Weiss, 1995). The information functions for all items can be summed to create a test information function (TIF) for the entire test (Reise & Henson, 2003). The TIF can also be used to compute a standard error of measurement that varies for different levels of θ:

$$SEM(\theta) = \frac{1}{\sqrt{(TIF(\theta))}} \tag{9}$$

(Reise & Henson, 2003).

IRT models have been developed for polytomous responses (see Embretson & Reise, 2000, for more detail). Once scales with polytomous responses have been modeled with IRT, then the IIC's, TIF's, and SEM's can be computed and interpreted just as for dichotomous models (Embretson & Reise, 2000).

Advantages and Disadvantages of IRT Models for Noncognitive Measures

Advantages of IRT

The current consensus concerning the IRT paradigm is that it is generally theoretically superior to the CTT paradigm for both cognitive and noncognitive psychometric test development in at least several respects (Chernyshenko, Stark, Chan,

Drasgow, & Williams, 2001; Embretson & Hershberger, 1999; Hays, Morales, & Reise, 2000; Reise & Henson, 2003; Rouse et al., 1999; Teresi, 2006; Waller et al., 1996; Weiss, 1995; Wilson et al., 2006a, 2006b): (1) The item parameters are invariant up to a linear transformation (see Rupp and Zumbo, 2006, for more details), that is, they are not dependent on the average level of the latent construct of the sample used to calibrate the item parameters. (2) The item and test parameters, i.e., the ICC parameters, and the IIC, TIF and SEM functions, can provide information about the quality of the items and the overall test across the entire range of the latent construct. Furthermore, the item and test information functions and the standard error are specific for each level of θ. Such information can be useful in constructing tailored tests that are more informative at certain levels of θ. (3) The fit of each item's ICC to the data can be tested. (4) Scoring in IRT is more precise because the parameters of the ICCs are used to weight the contribution of each item to the final score. (5) Item parameters and individual θ scores can be evaluated on the same scale.

Disadvantages of IRT

The main disadvantages of the IRT paradigm appear to center mostly around more pragmatic rather than theoretical concerns involving its proper implementation: (1) IRT requires a large number of subjects, preferably heterogeneous in nature on characteristics that are relevant to the latent construct(s) of interest, in order to accurately estimate the ICC parameters (Embretson & Reise, 2000). The exact figure for the minimum number of subjects is debatable, but at a bare minimum appears to be anywhere from 250 to 500 (Embretson & Reise, 2000). (2) The mathematical framework of IRT is far more intricate than the mathematical framework of CTT. Thus, the development

and/or analysis of a test in the IRT paradigm requires both the resources to collect a large number of observations and the possession of technical skills that are more advanced than those required for CTT.

Issues and Concerns Specific to the Application of IRT to Noncognitive Measures

Deciding to Use IRT for Noncognitive Constructs

The key issue in the decision to use the IRT paradigm for modeling noncognitive constructs is whether the increased theoretical and computational complexity of IRT is actually worth it (Reise & Henson, 2003; Waller et al., 1996). Reise and Henson (2003) posit three broad questions that need to be evaluated during the decision of whether to use the IRT paradigm in constructing and/or analyzing noncognitive constructs and their measures: "[1] Does IRT Significantly Change the Psychometric View of a Measure?… [2] Does IRT Make a Difference in Terms of Precision and Validity?… [3] Are IRT Models Appropriate for Personality Constructs" (pp. 99-100)? We will examine each of these questions in turn.

1. The Conception of Measurement in the IRT Paradigm Applied to Noncognitive Constructs.

Reise and Henson (2003) point out that many excellent noncognitive measures with high reliability and validity have been created under CTT. Furthermore, the ICC *a* and *b* parameters can actually be derived from CTT parameters with a reasonable degree of accuracy (Crocker & Algina, 1986; Reise & Henson, 2003). Furthermore, Wilson et al. (2006a) note that the procedures for determining the external validity of a measure do not change with IRT.

However, the property of invariance and the ability to model reliability for different levels of θ makes possible certain applications, such as test construction that targets specific levels of θ, the creation of shorter and more efficient tests, differential

item functioning, test equating, and computer adaptive tests, that are either not possible or much more difficult and less elegant under CTT (Reise & Henson, 2000, 2003; Rouse et al., 1999).

As an example of the power of IRT in helping to improve the conception of a noncognitive measure, Reise and Henson (2003) cited Gray-Little, Williams and Hancock (1997). Gray-Little et al. investigated the CTT and IRT properties of the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965, as cited by Gray-Little et al., 1997), which is a commonly used measure of self-esteem. Gray-Little et al. found that the RSE had overall excellent properties under both the CTT and IRT frameworks, however, the TIF revealed that the RSE's ability to accurately measure levels of self-esteem at θ values greater than +1 markedly decreased. Reise and Henson concluded that the decrease in test information for the RSE at higher θ scores "is a critical fact to know if a researcher were planning to use this measure to study change in self-esteem or trying to distinguish between people who, on average, have high self-esteem" (p. 99). An examination of Gray-Little et al.'s original article reveals that, curiously, the authors completely missed this important interpretation of their findings.

Waller et al. (1996) provided another good example of the power of IRT to assist in the creation of better noncognitive measures. As noted above, Waller et al. argue that the linear CTT models are inappropriate for noncognitive measures of the kind used in personality and clinical psychology. Dichotomous and polytomous Likert items have fundamentally a nonlinear relationship with the latent construct and using linear CTT models can lead to measures that are not as efficient as possible under IRT. To prove their point, Waller et al. created a 30 item Negative Emotionality Scale (NEM) from 122

out of 300 items of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982, as cited by Waller et al., 1996) that previous research had identified as being indicative of a higher order negative emotionality construct. Their brief version of the 30 item NEM possessed excellent discriminatory power across the full range of the latent construct and high external validity. Waller et al. conclude that it would not have been possible to create such a good brief 30 item scale from the 122 MPQ items without the framework of IRT.

2. Scoring, Reliability and Validity of Noncognitive Constructs in the IRT Paradigm

Reise and Henson (2003), as well as DeVellis (2006), point out that from personal experience they have often observed CTT and IRT methodologies produce scores for various tests that are highly correlated with each other. Using a dataset of scores from a large standardized aptitude assessment, Fan (1998) found that CTT and IRT parameters calibrated on multiple random subsamples from the large dataset consistently had a high correlation with each other. Thus, it would appear that for many common psychometric applications, the extra added complexity of IRT parameter calibration and scoring probably is not justified, especially if the only goal is to assign a set of scores.

However, Reise and Henson (2003) argue that even though the scores produced by CTT and IRT methodologies are highly correlated, there can be problems in using CTT based scores in behavioral research. For instance, as noted by Reise and Henson, Embretson (1996) found that the use of CTT based scores, but not IRT based scores, as outcomes in two-way experimental ANOVA designs can produce either spurious significant interactions or can mask true significant interactions. Also as noted by Reise and Henson, Fraley, Waller and Brennan (2000) provide evidence that the longitudinal growth curves of psychological test scores that originate at baseline from levels of θ

associated with low information can show spurious change over time as compared to growth curves of psychological test scores that have their origin at baseline from levels of θ associated with high information.

Reise and Haviland (2005) show that another advantage of IRT is that it can assist in judging clinically significant change from baseline in a longitudinal research paradigm. If a psychological test is used as a longitudinal outcome measure, the TIF at baseline can be used to create a "confidence band" (p. 234) that demarcates the zone of no significant change for each level of θ. Clinically significant change is defined to occur for an individual when his or her θ score goes beyond the zone of no change at subsequent time periods past baseline.

3. Appropriateness of the IRT Paradigm for Noncognitive Measures

> Reise and Henson (2003) have observed that they
>
> frequently encounter among assessment professionals and research colleagues [a misconception]. Namely, some researchers believe cognitive constructs are real, individual differences, psychobiological traits that cause behavior, whereas personality constructs are thought of as arbitrary, subjective, and merely summary labels of behavior. To many, it is thought that IRT methods are appropriate to use with cognitive variables but inappropriate to use with personality assessments. (p. 100)

Reise and Henson do agree that there are many "poorly thought out, redundant, intellectually flabby constructs and measures in personality assessment research" (p. 100), however they

> disagree with the view that personality measurement is a qualitatively different world than cognitive assessment. In many circumstances,

> personality constructs are deeply embedded within psychobiological theories and are properly viewed as real traits that cause behavior in the exact same way as cognitive variables like math ability or spatial ability. (p. 100)

In order to judge the appropriateness of applying IRT models to noncognitive data, Reise and Henson (2003) take the perspective, which is shared by a number of researchers in the field of noncognitive assessment (Crowley & Fan, 1997; Finch & West, 1997; Glockner-Rist & Hoijtink, 2003; McDonald, 1999; Panter, Swygert, Grant Dahlstrom, & Tanaka, 1997), that IRT models are conceptually equivalent to factor analytic models and that any noncognitive construct that can be modeled using factor analysis can also be modeled using IRT. However, Reise and Henson note that there are certain kinds of noncognitive constructs that are poor candidates for IRT modeling because they do not easily fit into the standard factor analytic framework: "multifaceted personality constructs" (p. 101), "nonlinear developmental constructs" (p. 101), and "emergent constructs" (p. 101).

Multifaceted personality constructs. Multifaceted personality constructs, which "are composed of multiple specific subcomponents" (Hull, Lehn, & Tedlie, 1991, p. 932), are resistant to accurate IRT modeling because it is difficult to satisfy the assumption of unidimensionality with such constructs. However, a potentially simple workaround in such situations is to model each subcomponent separately or to create models with second order latent factors/variables (Muthen & Muthen, 1998-2006).

Nonlinear developmental constructs. Developmental constructs are characterized by "milestone sequences" (Loevinger, 1993, p. 2), which are "age- or stage-specific

characteristics" (Loevinger, 1993, p. 2) that appear at a certain point during development and then disappear. Thus, the conceptualization of the latent scale of a developmental construct may be different from a nondevelopmental construct in that lower values of the construct correspond to earlier time points and/or stages of development and higher values of the construct correspond to later time points and/or stages of development (Loevinger, 1993; Noel, 1999). For such constructs, certain observed indicators may have a nonmonotonic relationship with the construct since they appear and disappear for specific time points and/or stages (Noel, 1999). Therefore, developmental constructs may violate the assumption of monotonicity for standard IRT models. However, new forms of IRT models, ideal point models (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Noel, 1999), may be able to handle developmental constructs. A detailed description of ideal point models is beyond the scope of the present review, but, briefly, they may be described as IRT models in which the ICC does not asymptote toward unity but rather it eventually decreases back down to the null value as levels of θ increase (Chernyshenko et al., 2007; Noel, 1999).

Emergent constructs. Emergent constructs are defined by Bollen and Lennox (1991) as constructs whose indicators cause the construct, i.e., an emergent construct is a linear combination of its indicator variables. In contrast to emergent constructs, traditional latent constructs have indicator variables that are influenced by the latent construct, i.e., a latent construct is theorized to be a causal influence on the levels of its indicator variables. A classic example of an emergent construct given by Bollen and Lennox is socioeconomic status (SES). According to Bollen and Lennox, SES actually makes for a bad traditional latent construct since different potential indicators of SES

found in the literature such as income, quality of neighborhood, and level of education essentially determine SES. As a result, SES should not be conceptualized as a latent variable in a factor analytic, structural equation modeling or IRT paradigm.

Concluding Thoughts on the Decision to Use IRT for Noncognitive Constructs

It appears that there are no clear black and white answers to the three questions posed by Reise and Henson (2003) concerning the use of the IRT paradigm for noncognitive measures. It does appear though that IRT can be enormously useful if researchers desire to construct a new test, or evaluate an existing test, that will be used in demanding research situations, such as using test scores to track individuals longitudinally or using test scores as outcomes in general linear models. However, even if the IRT paradigm is judged to be conceptually appropriate for a noncognitive measure, there are still critical concerns about choosing, implementing, and interpreting an IRT model. I now turn to examine some of these issues.

The Fit of IRT Models

A key concern for IRT modeling of latent constructs is whether IRT models accurately describe the relationship between the observed indicator variables and the underlying latent construct. Chernyshenko et al. (2001) observed that it has been shown through repeated empirical experience that logistic IRT models can be expected for the most part to show good fit in modeling the responses of individuals to cognitive test items. However, Chernyshenko et al. argued that there has not yet been enough empirical experience with using IRT to model noncognitive constructs to have the same level of confidence in the use of IRT for noncognitive measures as for cognitive measures.

Chernyshenko et al. (2001) explicitly tested the degree to which both parametric and nonparametric[2] IRT models fit personality data. They used 170 items from the 16 scales from the fifth edition of the 16 Personality Factor Questionnaire (16 PF; Conn & Rieke, 1994, as cited by Chernyshenko et al., 2001) and 50 items from the five scales of Goldberg's (1997, 1998, as cited by Chernyshenko et al., 2001) public domain Big Five personality instrument. The items from the 16PF have three response options: agree, disagree or don't know. The items from the Big Five have five Likert response options: from very inaccurate to very accurate.

Chernyshenko et al. (2001) investigated a set of parametric dichotomous (2PL, 3PL) and polytomous (Samejima's Graded Response Model [SGR]) IRT models as well as nonparametric dichotomous (Levine's Maximum Likelihood Formula Scoring [MFS]) and polytomous (Levine's polytomous MFS) IRT models (see Chernyshenko et al. for more details about the different models). For fitting the dichotomous IRT models, the response options for both tests were collapsed. For the 16PF, the don't know option was collapsed into the agree option. For the Big Five, the first three Likert response options were collapsed together and the last two Likert response options were collapsed together.

The IRT models were fitted after the items in each scale where shown to be unidimensional using modified parallel analysis and confirmatory factor analysis. To assess the fit of the IRT models, Chernyshenko et al. (2001) used modified graphical fit plots and adjusted chi-square goodness of fit techniques that incorporated cross-validation sampling techniques (see Chernyshenko et al. for more details). The chi-

[2] Nonparametric IRT models consist of ICC's that do not have a fixed rigid form as found in the 1PL, 2PL, and 3PL models, but instead have, as their name implies, ICC's that have a flexible nonlinear form that can accommodate a wide variety of item response patterns (see Junker, 2001; Mokken, 1971, 1997; Ramsay, 1991, 1997; Sijtsma & Molennar, 2002, for more information on nonparametric IRT models).

square tests of fit evaluated the fit of single items, all possible pairs of items within a scale, and all possible triplets of items within a scale. Chernyshenko et al. showed that the use of the chi-square goodness of fit test on pairs and triplets of items provides a stronger test of the fit of the ICC's than simply testing each item individually.

The findings of Chernyshenko et al. (2001) were somewhat complicated, and due to space limitations, Chernyshenko et al. were only able to report a representative subset of their results. For the 16PF, single item goodness of fit tests for all IRT models and scales were excellent and comparable to results previously found by members of the same research group for cognitive constructs (Drasgow, Levine, Tsien, Williams, & Mead, 1995). However, the results for the double and triple item goodness of fit tests showed a different picture. First, the SGR model showed poor fit for all scales. Second, one set of scales showed good fit for the 2PL and 3PL models, and for the dichotomous and polytomous MFS models. Third, a second set of scales showed poor fit for the 2PL and 3PL models, and a better fit for the dichotomous and polytomous MFS models. The graphical fit plots were examined in an attempt to make sense of the results of the goodness of fit tests. One striking observation was that the middle option for most 16PF items was relatively unused compared to the other two options, which most likely led to the poor fit of the SGR. Also, certain dichotomously scored items that had poor fit showed extremely unusual "V" shaped ICC's in the graphical fit plots.

For the Big Five, the single, double and triple item goodness of fit tests were poor across all scales for all the parametric models. The polytomous MFS models was not fitted due to a less than optimal sample size, but the dichotomous MFS model was able to be fitted and showed excellent fit for all scales.

Chernyshenko et al. (2001) concluded that "the issue of fitting IRT models to personality data is more complicated then previously suggested" (p. 554) and that the fit of IRT logistic ICC's should be carefully investigated when used for noncognitive items. Chernyshenko et al. attempted to determine whether a particular kind of item, e.g., positively keyed versus negatively keyed, had a tendency to show misfit. Their efforts were ultimately fruitless, except for discovering that there may have been some violation of the assumption of local independence for certain items because of the presence of some small level of multidimensionality. Chernyshenko et al. speculated that perhaps one of the reasons the parametric IRT models did not show uniform adequate fit across all scales of the 16PF and the Big Five tests is that the underlying processes involved in answering cognitive and noncognitive items may be different. Chernyshenko et al. drew on Cronbach's (1984) model of how individuals respond to different types of items as a possible explanation of their results.

Cronbach (1984) theorized that cognitive items attempt to ascertain "maximum performance" (p. 28) while noncognitive items attempt to ascertain "typical response" (p. 28). Items in the former case are useful "when we wish to know how well the person performs when asked to do his best" (Cronbach, 1984, p. 28), while items in the latter case "seek to appraise … what the person most often does or feels – in a recurring specific situation or in a broad class of situations" (Cronbach, 1984, p. 28). Chernyshenko et al. (2001) hypothesized that IRT may be able to more easily model tests that assess maximum performance as opposed to tests that assess typical response. For tests that model typical response, Chernyshenko et al. recommended exploring other

types of IRT models that were not initially designed for cognitive measures, such as the ideal point model (Chernyshenko et al., 2007).

Maydeu-Olivares (2005) replicated in part Chernyshenko et al.'s (2001) study using five subscales of the Social Problem Solving Inventory – Revised (SPSI-R; D'Zurilla, Nezu, & Maydeu-Olivares, 2002, as cited by Maydeu-Olivares, 2005) and came to somewhat more optimistic conclusions concerning the application of standard IRT models to noncognitive measures. The SPSI-R has five Likert response options and was designed explicitly to have unidimensional scales. Maydeu-Olivares fitted four different polytomous parametric IRT models and Levine's nonparametric polytomous MFS model. Maydeu-Olivares hypothesized that since the SPSI-R scales were designed to be unidimensional, parametric IRT models should fit well. Furthermore, in contrast to Chernyshenko et al.'s conjecture that IRT models may not be appropriate for noncognitive measures because of underlying typical response processes, Maydeu-Olivares hypothesized that "ideal point models … may be more appropriate for some attitude data. But for personality data, where often respondents are asked the degree with which a description applies to them, or how often they perform certain behaviors, [IRT models] should be … appropriate" (p. 266).

Maydeu-Olivares' (2005) results were complicated and their summary here will necessarily be brief and only focus on general patterns of findings. Maydeu-Olivares used goodness of fit tests that were identical to Chernyshenko et al (2001). It was found that at least one of the parametric models outperformed the nonparametric model for every scale of the SPSI-R as gauged by the goodness of fit tests on single items in the calibrating sample. The nonparametric model tended to outperform all parametric

models as gauged by the goodness of fit tests for pairs and triplets of items in the calibrating sample. For the cross-validating sample, Maydeu-Olivares found a mixture of results in that the goodness of fit tests did not show any overwhelming evidence of the overall superiority of either the parametric or nonparametric models in modeling noncognitive scales. Maydeu-Olivares concluded that parametric IRT models can indeed be used to model noncognitive items since their rates of successful cross validation were comparable to nonparametric IRT models for a test designed to be unidimensional.

The Meaning of the c parameter in 3PL IRT Models

3PL IRT models, which have a *c*, or guessing, parameter were developed to handle the possibility of guessing on cognitive items (Embretson & Reise, 2000; Hambleton et al., 1991). The meaning of the *c* parameter for noncognitive items is less clear since test takers are not expected to guess on their responses for attitudinal, personality, or clinical measures (Rouse et al., 1999). For this reason, some researchers have tended to shy away from using the 3PL model with noncognitive measures (Chernyshenko et al., 2001). Furthermore, some investigations into the use of 3PL models for noncognitive measures have found no substantial improvement in fit as compared to other IRT models (Chernyshenko et al., 2001). However, other studies suggest that the 3PL IRT model may indeed be useful for noncognitive measures (Ellis, Becker, & Kimmel, 1993; Reise & Henson, 2003; Rouse et al., 1999; Zumbo, Pope, Watson, & Hubley, 1997). If the 3PL model is useful in modeling noncognitive measures, then the question of what the *c* parameter means for noncognitive measures must be considered.

Zumbo et al. (1997) hypothesized that the *c* parameter for noncognitive measures may reflect "a social desirability bias" (p. 963), i.e., if the low end of the latent spectrum of a noncognitive trait is viewed as undesirable by test takers then the test takers may have an increased propensity to positively respond to items that are keyed toward the high end of the latent trait. Rouse et al. (1999) tested Zumbo et al.'s hypotheses using the Personality Psychopathology Five (PSY-5; Harkness, McNulty, & Ben-Porath, 1995) test, which uses items from the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 1989, as cited by Rouse et al., 1999) to measure five personality dimensions: (1) Aggressiveness, (2) Psychoticism, (3) Constraint, (4) Negative Emotionality/Neuroticism (NEM), and (5) Positive Emotionality/Extraversion (PEM). These five personality dimensions represent "five key pieces of information that one would want to know about another person in many interpersonal situations, ranging from understanding the personality of a potential roommate to summarizing the personality pathology of a psychiatric inpatient" (Rouse et al., 1999, p. 293). Rouse et al. successfully fitted a 3PL model to all five scales of the PSY-5 and found the following ranges for the *c* value for each scale: Aggressiveness (.01-.08), Psychoticism (.06-.36), Constraint (.03-.22), NEM (.01-.07), and PEM (.02-.26). It appears that individuals tend to have a bias toward representing themselves as high on Psychoticism, Constraint, and PEM. Rouse et al. then correlated the *c* parameters with the social desirability (SoD) ratings that were available for each MMPI-2 item and found substantial correlations between the SoD ratings and the Aggressiveness (.49) and Psychoticism scales (.60), as well as a moderate correlation between the SoD ratings and the NEM scale (.31). Rouse et al. concluded that "the three scales that showed the strongest relations between SoD

and the *c* parameter were three scales that could be conceptualized as unidirectionally pathological, namely, Aggressiveness, Psychoticism, and NEM" (p. 303). These findings are difficult to interpret because two of the three highest correlations between *c* and the SoD ratings were for scales (Aggresiveness and NEM) that appeared to measure socially undesirable traits, at least as gauged by their relatively low *c* parameter values. Most likely, the data collected by Rouse et al. requires a more indepth statistical and content analysis of the scales and their individual items to make more sense of their results. Rouse et al. call for more future research to explore the meaning of the *c* parameter for noncognitive items.

Reise and Waller (2003) conducted a comprehensive comparison of the use of 2PL versus 3PL IRT models on a set of noncognitive measures with the goal of acquiring a greater understanding of what the *c* parameter means for noncognitive items. For their investigation, they used 15 scales from the adolescent version of the MMPI (MMPI-A; Butcher et al., 1991, as cited by Reise & Waller, 2003) that were extracted using factor analysis on a matrix of tetrachoric correlations of the 15 scales. They fitted a 2PL model and a 3PL model for each scale in the standard keyed direction, which in this case means that item responses reflected the presence of psychopathology. In addition, they also fitted a 3PL model in the reversed keyed direction (3PL-R), which means that the item responses where flipped so that they reflected the absence of psychopathology.

Reise and Waller's (2003) rational for scoring the scales in both the keyed and reversed keyed directions was that most noncognitive items, unlike cognitive items, do not have an obvious direction for keying, and therefore different kinds of information may be present at both ends of the latent spectrum. For instance, in a test of

mathematical aptitude, it is obvious that the keying of the item should be in the direction of greater mathematical aptitude since it makes no sense for interpretative purposes to determine the degree to which an individual lacks mathematical ability. However, for an extroversion scale, there is some arbitrariness as to whether an item should be keyed in either the extroversion or nonextroversion/introversion direction. Reise and Waller also hypothesized that that by "estimating the 3PL under both directions of scale keying, … [they hoped] to more fully capture and illustrate the extent to which personality items fail to conform to the [2PL model]" (p. 167).

Reise and Waller (2003) used two methods to determine whether the standard and reversed keyed 3PL models significantly differed from the 2PL model for each scale. Their first method was a chi-square deviance test, in which the differences in the -2 log likelihood statistics between the 2PL model and both kinds of the 3PL models were tested with a chi-square test. The second method involved comparing the root mean square residuals (RMSR) between the 2PL model and the 3PL models across all items for a given scale. For most scales, either one or both types of the 3PL models was significantly different from the 2PL model as determined by the deviance test, which indicated a better fit of the 3PL model and/or the 3PL-R model over the 2PL model. In contrast, the RMSR statistics did not appreciably differ across the three kinds of IRT models.

The resolution of these seemingly paradoxical results was found in the distribution of the *a* and *c* parameters in the different IRT models. The average value of the 3PL *c* parameter was close to zero across most scales, while the average value of the 3PL-R *c* parameter was around .10 across most scales. In addition, across most scales the

distribution of the 3PL-R *c* parameter had a greater degree of variability than the 3PL *c* parameter and a greater degree of skewness, with a large number of observations above .10. The average value of the *a* parameter was lower in the 2PL model than in either one or both types of the 3PL models across most scales.

According to Reise and Waller (2003), the distribution of the *a* and *c* parameters shows that the 2PL model compensated for the lack of a nonzero asymptote by lowering the *a* value as compared to its 3PL counterparts. In the 3PL models, the presence of a nonzero asymptotic *c* parameter allowed the *a* parameter to be higher, i.e., the ICC was able to become more discriminating because it did not have to stretch out its curve to reach an asymptote at zero. Reise and Waller argued that while either one or both of the 3PL IRT models fit better as judged by the -2 log likelihood tests than the 2PL model across most scales, the 2PL model compensated for the lack of a nonzero asymptote by decreasing its discriminability across items for most scales, which ultimately lead to RMSR statistics that did not differ across the different models.

From a substantive diagnostic perspective, the results of Reise and Waller (2003) show that, for many of the psychopathology scales, individuals who had high latent θ scores did not respond in the expected psychopathology direction for approximately a third of the items. The opposite pattern, i.e., individuals with low θ scores affirmatively responding to items in the psychopathology direction, was also observed, but to a far lesser extent, i.e., only about a tenth of the items. From a mathematical perspective, the results of Reise and Waller indicate that essentially certain items do not have an upper asymptote at unity as is assumed by the standard 3PL IRT model, and this phenomena is only detected with 3PL IRT models by reverse keying all items.

To further investigate the meaning of the *c* parameter, Reise and Waller (2003) conducted a content examination of items that had *c* parameter values greater .10. They concluded that for the most part such items had some amount of ambiguity in their content at one end of the latent spectrum. "In other words, at one end of the continuum, the item is a good marker of a single latent trait, whereas at the opposite end the item is related to several traits" (Reise & Waller, 2003, p. 175). Thus, in contrast to Rouse et al. (1999), Reise and Waller concluded that their results demonstrated that a large *c* parameter for noncognitive data is not necessarily the result of some sort of deception related to socially desirable responding or "faking good" (p. 181) on the part of the test takers.

Reise and Waller (2003) concluded that for researchers to acquire a deeper understanding of any noncognitive scale they should examine the scale's items in an IRT framework for the possibility of nonunity upper asymptotes. A 4PL model (see Barton & Lord, 1981, as cited by Reise & Waller, 2003, and Hambleton & Swaminathan, 1985, p. 49, as cited by Reise & Waller, 2003) that allowed for a nonzero lower asymptote and a nonunity upper asymptote would be useful to use in such cases, however, no commercially available software program has the capability to fit such a model (Meijer & Baneke, 2004). However, Reise and Waller argued that an acceptable and convenient alternative to the 4PL model is to fit both standard and reversed keyed items to 3PL models as they have done. In addition, researchers should pay closer attention to the content of the items in any noncognitive scale that they create and/or evaluate. The use of such a set of procedures would allow researchers to acquire a greater understanding of how individuals respond to noncognitive items.

Conclusions Concerning the Applicability of IRT Models to Noncognitive Constructs

It appears that IRT models can indeed be used to model noncognitive constructs. As seen above, though, there can be challenges in applying IRT models to noncognitive constructs. However, these potential concerns do not appear to be relevant for the IRT modeling of psychodiagnostic symptom criteria, such as those of depression.

One primary concern about the applicability of IRT models to noncognitive constructs as reviewed by Reise and Henson (2003) was that noncognitive constructs may not be theoretically robust enough. Psychodiagnostic symptom criteria do not seem to have this problem because they are embedded in a strong multilevel theoretical framework that spans from the micro level of individual bio-cellular processes to the macro level of broad socio-historical contexts (Barlow & Durand, 2005). Another concern voiced by Reise and Henson about applying IRT models to noncognitive constructs is whether they are appropriately modeled by techniques for latent constructs such as factor analysis. For Reise and Henson, this is only a concern if the construct in question is multifaceted, developmental, or emergent in nature. This does not appear to be an issue for psychodiagnostic constructs such as depression for reasons that will be outlined below.

Psychodiagnostic symptom criteria may be multifaceted for a particular disorder, however, as noted above, this can easily be handled by treating subcomponents of a hypothesized psychopathological construct separately. Psychodiagnostic symptom criteria are not developmental constructs in the sense that they are not typically expected to appear and disappear as a matter of normal development (Barlow & Durand, 2005; Loevinger, 1993). Of course, it is well known that many psychopathological symptoms

appear during adolescence and/or negative life changing events, and that psychological and/or biological interventions, as well as positive life changing events, may reduce or eliminate symptoms (Barlow & Durand, 2005). However, the appearance and disappearance of psychopathological symptoms is not specifically intrinsic to normal developmental processes that follow an orderly expected pattern of development. In fact, many psychopathological problems may be considered a derangement of normal biological, cognitive and/or social developmental processes (Barlow & Durand, 2005). Therefore, psychodiagnostic symptom criteria should not be considered developmental from a psychometric perspective. Psychodiagnostic criteria are not emergent constructs, since most paradigms of psychopathology assume that psychopathological symptoms are actually caused by underlying cognitive-neurological and/or global socio-historical processes (Barlow & Durand, 2005).

Chernyshenko et al. (2001) have a concern that IRT models may not be applicable to noncognitive constructs because noncognitive tests may assess what Cronbach (1984) labeled as "typical response" (p. 28) rather than "maximum performance" (p. 28). However, a careful content analysis of many of the symptom criteria for disorders such as depression and anxiety in DSM-IV shows that the symptom criteria do indeed assess maximum performance in that individuals who are truly in a state of mental and emotional distress can only affirmatively respond to those symptoms (APA, 2000). Once an individual passes a certain threshold of mental and emotional distress, there is no possible way that they can negatively respond to most of the DSM-IV symptom criteria, which would be the case if the symptom criteria were assessing typical response (APA,

2000). Therefore, it appears that IRT models do have the strong theoretical potential to have a good fit to the individual symptom criteria for depression.

Overall, IRT can provide a sophisticated analysis of psychodiagnostic symptom criteria for a disorder such as depression. If a 1PL model is found to be plausible for a set of symptom criteria, then the use of a straightforward total symptom count during the diagnostic procedure is justified (Embretson & Reise, 2000; Hambleton et al., 1991; Reiser, 1989). If a 2PL model holds true, then use of the raw total score may be problematic (Embretson & Reise, 2000; Hambleton et al., 1991). If a 3PL model holds true, then there would be evidence that perhaps there is a tendency on the part of individuals and/or diagnosing clinicians to over report certain symptoms at low levels of a particular disorder or even in the absence of any psychopathology (Rouse et al., 1999). On the other hand, if a 3PL-R model holds true, then there would be evidence that certain symptoms are underreported for individuals who have high levels of a particular disorder (Reise and Waller, 2003).

Depression

Construct of Depression and its Measurement

Angst and Merikangas (1997) observed that during the past 50 years, research on depression has seen "rapid progress" (p. 31) due to several factors, which include the development of better treatments and the use of longitudinal prospective research designs. Central to all the current research on depression is the construct of depression itself. As a psychological construct, depression is not of course directly observable, however there have been many attempts to scientifically define it throughout history (Frances et al., 1990; Kendler, 1990). As reviewed above, the most current "official"

definition of depression for North America is the one found in DSM-IV TR, which consists of a nine symptom checklist, as reviewed above, with a cutpoint of five symptoms required for a diagnosis, one of which needs to be depressed mood or anhedonia (APA, 2000). After many years of various definitions of depression, many of which were concurrently operative in different geographical and clinical contexts, "this sort of operational definition, progressively refined ... has been revolutionary" (Kramer, 2005, p. 159). Kramer (2005) concluded that

> it is impossible to overstate the influence or the success of the operational definition. It has been a more important scientific tool than the PET scan. Almost every research result regarding depression in humans refers to people with at least two weeks of five symptoms of moderate severity. The altered neuroanatomy, the genetic risk, the excess disability – all are liabilities of major depression, operationally defined. (p. 160)

However, the current DSM definition of depression is rife with controversy (Barlow & Durand, 2005; Clark et al., 1995). One of the most controversial aspects of the DSM definition of depression, as with several types of disorders defined by the DSM such as the anxiety disorders, is whether depression should be construed as a categorical or continuous construct (Carroll, 1984; Krueger & Piasecki, 2002). Because of the way it is structured, the DSM-IV definition of depression is explicitly categorical, i.e., an individual either is depressed or is not. Alternatively, some researchers have proposed that a continuous model of depression would be more adequate (Clark et al., 1995).

The creators of DSM-IV have acknowledged the limits of the categorical approach for psychiatric diagnoses, however, they noted that the "naming of categories is

the traditional method of organizing and transmitting information in everyday life and has been the fundamental approach used in all systems of medical diagnosis" (APA, 2000, p. xxxi) and thus, this approach was retained for the DSM-IV. The creators of DSM-IV also noted that while dimensional approaches toward psychopathology "increase reliability and communicate more clinical information (because they report clinical attributes that might be subthreshold in a categorical system)" (APA, 2000, p. xxxii), they are fraught with their own limitations, including unfamiliarity for psychiatric clinicians who are used to a more traditional medical model of diagnosis based on a categorical approach, a lack of evocative labels and descriptions for various psychopathologies, and a lack of consensus as to how many dimensions are needed for describing psychopathologies (APA, 2000). For these reasons, the creators of DSM-IV decided to retain the categorical approach toward the various kinds of psychopathologies.

A psychometric approach toward psychopathology can provide useful information and a deeper perspective for both sides of the categorical/continuous divide. For proponents of the categorical approach toward psychopathology, the psychometric perspective can show, for instance, what kind of information different criteria are providing toward a categorical diagnosis and how the information from different criteria can be efficiently combined for an optimal set of diagnostic algorithms. On the other hand, for proponents of the continuous approach toward psychopathology, the psychometric perspective can show, for instance, how well the different symptom criteria for a diagnostic construct cover the entire range of the continuous latent scale of the diagnostic construct.

Cutpoint requirement of five symptoms

A prominent feature of the DSM-IV diagnostic rules for an MDE is the cutpoint of five symptoms, one of which must be one of the two gate criteria, imposed on a total possible maximum of nine diagnostic symptom criteria. There has been much controversy surrounding the adequacy of this cutpoint criterion. A number of researchers have hypothesized that the presence of less than five symptoms can also be psychiatrically debilitating and, as a result, have created a diagnostic category known as subthreshold depression (Angst & Merikangas, 1997; Judd, Akiskal & Paulus, 1997; Sadek & Bona, 2000).

Subthreshold depression is defined as the presence of a certain number of DSM depression symptoms that do not meet certain restrictions or qualifiers found in the original DSM diagnosis of depression such as the minimum cutpoint of five symptoms or the presence of symptoms for two weeks. Several categories of subthreshold depression have been proposed, all of which are based on modifications of the DSM diagnostic rules for depression (Angst & Merikangas, 1997; APA, 2000; Judd et al., 1997; Kessler, Zhao, Blazer & Swartz, 1997; Maier, Gänsicke, & Weiffenbach, 1997; Sadek & Bona, 2000): *minor depression*, which is defined similar to a MDE except that the cutpoint for the number of symptoms is two to four, *subsyndromal symptomatic depression*, which is defined as the presence of two or more MDE symptom criteria without the presence of either one of the gate criteria of depressed mood or anhedonia, *recurrent brief depression*, which is defined as the presence of all required symptom criteria for a full blown MDE except that the symptoms last less than two weeks, and *dysthymia*, which is defined in DSM-IV as a persistent low mood, which is less intense than a full blown

MDE, that lasts for at least two years. Each of these subcategories of depression is hypothesized to reflect different intensities of depression at levels below a full blown episode of depression. Most of the research on these different subcategories of depression has found that they are associated with legitimate psychiatric distress and numerous concomitant psychosocial disabilities (Angst & Merikangas, 1997; APA, 2000; Judd et al., 1997; Kessler et al., 1997; Maier et al., 1997; Sadek & Bona, 2000).

Kessler et al. (1997) examined the implications of having different levels of symptom criteria of depression using data collected from the first National Comorbidity Survey in the early 1990's. Kessler et al. created three categories of depression: minor depression, which was defined by having 2-4 symptoms, MD 5-6, which was defined by having 5-6 symptoms, and MD 7-9, which was defined by having 7-9 symptoms. Extensive analyses were done on the differences between these three different categories of depression with regard to sociodemographic factors, course of illness, lifetime prevalence, and various indicators of life impairment such as days missed from work, subjective assessment of overall well being and success in life, visiting doctors, and use of depression medication. Kessler et al. found that

> there is a clear gradient of increasing impairment from [minor depression] to MD 7-9 for each of these indicators. A substantial minority of those with [minor depression] (42.0%) and larger proportions of those with MD 5-6 (49.7%) and MD 7-9 (68.2%) reported at least one of these indicators of impairment. The differences in impairment between [minor depression] and MD 5-6 are consistently as small as or smaller than those between MD 5-6 and MD7-9,

> implying that there is *not* [italics added] a sharp divide between the lifetime impairments associated with [minor depression] and [major depression]. (p. 24)

Kessler et al. (1997) also found in an analysis of the lifetime prevalence of minor depression, MD 5-6 and MD 7-9 that the recurrence rates for all three types of depression are both high and similar (approximately in the low 70% range). There were also similar patterns of sociodemographic indicators across all three categories of depression. Based on their findings, Kessler et al. concluded that "[minor depression] cannot be dismissed as simply a normal reaction to environmental stress while [major depression] is seen as something quite different … [minor depression] cannot be dismissed as merely a transient mood state while [major depression] is seen as a chronic condition" (p. 28).

Sadek and Bona (2000) reviewed a number of studies that correlated various subthreshold depressive categories with psychosocial impairment. From a psychometric perspective, psychosocial impairment can be considered an external validating criterion of the construct of depression in that higher levels of depression are hypothesized to be associated with greater psychosocial impairment. Sadek and Bona concluded in a review of the available literature on the relationship between different subthreshold categories of depression and psychosocial impairment that "psychosocial impairment can indeed result from mild symptoms of depression, which even do not satisfy DSM-IV criteria for any depressive category" (p. 36).

Kendler and Gardner (1998) conducted an ambitious study in which they attempted to directly test the validity of the cutpoint of five symptoms by using regression techniques to determine whether "major depression [is] a discrete syndrome with 'points of rarity' at its boundaries[.] That is, is there a discontinuity in etiologic

processes so that major depression differs qualitatively and not just quantitatively from subsyndromal conditions" (p. 172)? Kendler and Gardner found that a straightforward linear regression function predicted future episodes of depression as a function of the number of previous depressive symptom criteria with no statistically significant discontinuity between four and five symptoms. Thus, they concluded that their results suggest that the cupoint of five symptom criteria for a diagnosis of depression does not "appear to carve nature at its joints" (p. 176) and that the "current DSM-IV diagnostic conventions for major depression … may be arbitrary and not reflective of a natural discontinuity in depressive symptoms as experienced in the general population" (p. 177).

The above reviewed literature on subthreshold depression has some important implications for the treatment of depression in that the best treatment outcome for depression appears to be complete elimination of all symptoms. As Kramer (2005) notes, "Even modest disruptions of sleep and appetite, for example, signal a substantial increased likelihood of future episodes and all they imply in terms of harm. By the late 1990's, it had become clear that *symptom-free* [italics added] recovery is the goal in the treatment of depression" (p. 164).

Psychometric Modeling of the Diagnostic Symptom Criteria of Depression

CTT Modeling of DSM Depression Criteria

There have been a small number of studies that have used CTT to assess the psychiatric diagnostic symptom criteria for depression.[3] The goal of all these studies was

[3] Note that this review does not include studies that have used psychometric techniques to assess various depression instruments such as the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the Hamilton Rating Scale for Depression (Hamilton, 1967), and the Center for Disease Control Depression Scale (Radloff, 1977).

essentially to determine how much each of the symptoms of depression individually contributed to the diagnosis of depression.

Faravelli et al. (1996) carried out a relatively simple CTT analysis of the diagnostic symptoms of depression as defined in DSM-IV. Their subject pool was 196 patients in outpatient treatment in Italy. Faravelli et al. conducted an item analysis of the individual symptoms by computing the item difficulty for each symptom and the correlations between individual symptoms and the total number of symptoms. The DSM-IV symptoms of depressed mood (p = .99; ρ_{XT} = .60), anhedonia (p = .76; ρ_{XT} = .60), and guilt (p = .52; ρ_{XT} = .60) had the highest correlations with the total number of symptoms (Faravelli et al., 1996). The DSM-IV symptoms of impaired memory (p = .30; ρ_{XT} = .22), sleep problems (p = .31; ρ_{XT} = .15) and irritability (p = .04; ρ_{XT} = 0) had the lowest correlations with the total number of symptoms (Faravelli et al., 1996). Overall, Faravelli et al. found

> that the greater the severity of the symptoms, the higher the number of symptoms. In other words, having a single severe depressive symptom increases the probability of having more symptoms. However, the most typical depressive symptoms bore greater correlation to the total number of symptoms than the less typical, and this contrasts with the assumption of quantitative classifications that all symptoms have the same value. (p. 309)

Faravelli et al. concluded that their results show that certain symptoms contribute more toward the diagnosis of depression than others.

Another CTT study on the diagnostic symptoms of depression was done by Buchwald and Rudick-Davis (1993). Their subject pool was 168 patients in outpatient

treatment in Texas. Buchwald and Rudick-Davis assessed the symptoms of depression in DSM-III by computing the positive predictive value, the negative predictive value and the total predictive value for each symptom with regards to the dichotomous outcome of a diagnosis of depression. The positive, negative and total predictive values are statistical techniques from epidemiology that conceptually can be considered similar to item total score correlations in that they are a measure of the strength of the relationship between each symptom and a dichotomous diagnostic outcome measure. The positive predictive value for a particular symptom is the percent of individuals who have that particular symptom and are eventually diagnosed with the illness associated that symptom (Gordis, 2000; Meehl & Rosen, 1955, as cited by Buchwald & Rudick-Davis, 1993). The negative predictive value for a particular symptom is the percent of individuals who do not have that particular symptom and are not diagnosed with the illness associated with that symptom (Gordis, 2000; Meehl & Rosen, 1955, as cited by Buchwald & Rudick-Davis, 1993). The total predictive value for a particular symptom is the percent of individuals who are correctly diagnosed with or without an illness based on both the presence and absence of the symptom (Meehl & Rosen, 1955, as cited by Buchwald & Rudick-Davis, 1993). Buchwald and Rudick-Davis found that most of the symptoms of depression have fairly high range of positive predictive values (.76 - .92, with most values in the .80's) and had moderately high total predictive values (.68 - .85). With the two notable exceptions of the symptoms of sleep problems and a lack of energy, which had high negative predictive values of .94 and .84, respectively, most of the symptoms had moderate negative predictive values (.53 - .66) (Buchwald & Rudick-Davis, 1993). Buchwald and Rudick-Davis concluded that a careful analysis of the pattern of their

results suggested two main findings concerning the symptoms of depression. First, "there were no indications that there are two or more distinct syndromes among the cases of [major depressive disorder]" (p. 204). Second, two symptoms of depression, "loss of energy and thinking difficulties … have the largest differences between true-positive and false-positive rates … Thus, these two symptoms are most strongly associated with [major depressive disorder], by several criteria" (pp. 204-205).

Zimmerman et al. (2006a, 2006b) have carried out perhaps one of the most ambitious examinations of the diagnostic criteria of MDD using the CCT paradigm. Their subject pool was quite large, 1,523 patients in outpatient treatment in Rhode Island, with a demographic profile of 60.5% female and 87.1% Caucasian. The results of their studies were published in a series of 12 consecutive papers over the course of one year in the *Journal of Nervous and Mental Disease*. A summary of all their results is beyond the scope of the present review; however, the results of the first paper of the series (Zimmerman, 2006a) are relevant for a discussion of a CTT analysis of the diagnostic symptom criteria for depression.

Zimmerman et al. (2006a) investigated the nine DSM-IV symptom criteria of depression using a set of statistical indexes from epidemiology (sensitivity, specificity, the odds ratio, positive predictive value, and negative predictive value) that modeled the relationship between individual symptom criteria and the dichotomous outcome of a diagnosis of depression. The positive predictive and negative predictive values were defined above. The odds ratio is a well known statistic in categorical data modeling (see Agresti, 1996, for more detail). In epidemiology, the sensitivity and specificity of a diagnostic test are measures of its validity (Gordis, 2000). The sensitivity of a test "is

defined as the proportion of diseased people who were *correctly* [italics added] identified by the test" (Gordis, 2000, p. 64). Gordis (2000) gives the formula for sensitivity as:

$$Sensitivity = \frac{Number\ of\ True\ Positives}{(Number\ of\ True\ Positives + Number\ of\ False\ Negatives)}. \quad (10)$$

Conversely, the specificity of a diagnostic test "is defined as the proportion of nondiseased people who are *correctly* [italics added] identified as negative by the test" (Gordis, 2000, p. 65). Gordis (2000) gives the formula for specificity as:

$$Specificity = \frac{Number\ of\ True\ Negatives}{(Number\ of\ True\ Negatives + Number\ of\ False\ Positives)}. \quad (11)$$

In Zimmerman et al.'s analyses of the sensitivity and specificity of the symptoms of depression, each symptom criterion was treated as a "test" that predicted the diagnosis of depression.

> Zimmerman et al. (2006a) found that
>
> at the level of the DSM-IV criterion, sensitivity varied between 55% and 93%. Six of the nine DSM-IV criteria has sensitivities above 75%, and six of the nine had specificities above 75%. Three criteria, depressed mood, anhedonia, and worthlessness/guilt, achieved both a sensitivity and specificity above 75%.… The odds ratios of all criteria were significant.… All nine DSM-IV diagnostic criteria had positive predictive values above 75%. Both the depressed mood and anhedonia items had positive predictive values above 85%. (p. 161)

Zimmerman et al. (2006a) also created a multiple logistic regression where all symptoms of depression were predictors of a diagnosis of depression. They found that all

nine symptoms were significant predictors of a diagnosis of depression. The symptoms of depressed mood (b = 3.08) and anhedonia (b = 2.38) were the strongest predictors of a diagnosis of depression, while a decrease or increase in psychomotor activity (b = 1.08) and the presence of suicidal tendencies (b = .94) were the weakest predictors of a diagnosis of depression.

Zimmerman et al. (2006a) concluded that overall their results show that:

> First, there is a variability in the frequency of the diagnostic criteria/symptoms with insomnia, fatigue, and impaired concentration most frequent, and suicidality and psychomotor disturbance the least frequent symptoms.… Second, symptom sensitivity and specificity tended to be inversely related. (p. 163)

From a CTT perspective, the three studies of Faravelli et al. (1996), Buchwald and Rudick-Davis (1993), and Zimmerman et al. (2006a) had some serious limitations. First, and probably most serious, is that the dimensionality of the set of depression criteria was not assessed through factor analysis in any of the three studies. Second, the overall reliability of the diagnostic criteria using Cronbach's α as well as the SEM for the total score were not assessed in any of the studies. It would be useful to know the overall reliability of the total score since the diagnostic criteria for depression in all the nosologic systems that were examined had a minimum number of symptoms that had to be observed, i.e., a cutpoint, in order for a diagnosis of depression to be made. Third, two of the studies (Buchwald & Rudick-Davis, 1993; Zimmerman et al., 2006a) correlated individuals' symptoms with the dichotomous outcome of a diagnosis of depression; however, the underlying polythetic symptoms clearly fall on a continuum. Therefore, correlations should have been reported between the symptoms and the total count of

symptoms in those two studies. Fourth, each of the above three studies was based on data sets that were not representative of the entire population. Buchwald and Rudick-Davis and Faravelli et al. were based on relatively small datasets from clinical outpatient populations, and Zimmerman et al. was based on a dataset that had mostly female and white outpatients.

Overall, a careful review of the methodology, the interpretation of the results, and the conclusions of the above three studies appears to indicate that the authors did not have a good grasp of CTT. For instance, Faravelli et al. (1996) observed that the symptoms of depression that are less likely to occur have a weaker correlation with the total symptom count and they reported this as a major finding. However, it is actually well known in the CTT literature that items that have high or low difficulty values will have lower correlations with the item total score than items with moderate difficulty values (Crocker & Algina, 1986). Two of the studies reviewed above (Buchwald & Rudick-Davis, 1993; Faravelli et al., 1996) were not quite true CTT studies in the sense that they relied on statistical techniques developed in the field of epidemiology, e.g., sensitivity, specificity, positive predictive value and negative predictive value, and they did not rely enough on techniques that have been developed and refined in the CTT literature, e.g., factor analysis, item correlations with total score, and Cronbach's alpha (Crocker & Algina, 1986).

IRT Modeling of DSM Depression Criteria

A search of the literature revealed two prominent studies (Aggen et al., 2005; Reiser, 1989) that used IRT to assess the symptom criteria of depression. Both of these studies used only 1PL and 2PL IRT models.

Reiser (1989) attempted to fit a series of 1PL models to the DSM-III depression symptom criteria. Reiser used data from the Epidemiologic Catchment Area (ECA) survey, which sampled from five cities in the United States (Kessler & Üstün, 2004). Reiser ran his models on data mostly from one of the cities, Baltimore, in the ECA. He found that the 1PL model had a poor fit to the data, as judged by indices of fit based on the G^2 and χ^2 statistics. However, fit indices for the 1PL IRT model improved considerably if two different subsets of symptom criteria (appetite/eating problems, sleeping problems, sexual problems, and suicidal ruminations grouped as one subset and psychomotor changes, low energy, low self-esteem/guilt, and cognition problems grouped in another subset) were created in which each subset was constrained to have the same *b* value but the *b* values were allowed to be different between the two subsets.

Reiser (1989) also tested the fit of 2PL IRT models for an extensive variety of different subsets of the symptom criteria and found that the 2PL IRT models for subsets of symptoms that did not have the depressed mood criterion fit much better than if the depressed mood criterion was included. Reiser hypothesized that the problems with the depressed mood criterion may have been due to a lack of self-awareness of low mood in depressed individuals, which may may lead depressed individuals to negatively respond to this symptom. As evidence for his speculation, Reiser noted that the lack of self-awareness of low mood in depressed individuals has been clinically observed. Reiser concluded that based on his analyses, there may not be a single latent continuum underlying the criteria for depression.

Aggen et al. (2005) successfully fitted a 1PL and 2PL model to the DSM-IIIR depression symptom criteria on data from 2,163 Caucasian identical female twins that

were born in Virginia. Aggen et al. first determined the unidimensionality of the symptom criteria through exploratory and confirmatory factor analyses on tetrachoric correlations of the symptoms. The 2PL model fit better than the 1PL model at statistically significant levels as judged by -2 log likelihood statistic. The ICC's for the 2PL model had *a* parameter values between 1.67 and 3.21 and *b* parameter values between .35 and 2.29. About five of the symptom ICC's in the 2PL model (anhedonia, change in weight or appetite, sleep problems, change in psychomotor activity, and low energy) tended to cluster together on the latent continuum of depression, which indicated that they provided approximately the same amount of information for diagnostic purposes.

Aggen et al. (2005) also externally validated the IRT model of depression by creating two separate regression models in which scores on a measure of neuroticism and a later independent diagnosis of depression, respectively, where regressed onto the initial latent θ depression scores computed from the 1PL model, the binary diagnostic score of the presence of depression, age, and the interaction between the latent θ depression scores and the binary diagnostic score. They also created a second set of two separate validation regression models that were identical to the first set except that they used latent θ depression scores computed from the 2PL model. They found that the continuous latent θ depression score was a significant predictor in both kinds of models while the binary diagnostic score was an insignificant predictor. Based on Aggen et al.'s results, it does appear that using the continuous latent scale for depression offers more information than simply dichotomizing at a cut point of five or more symptom counts.

Both studies that used IRT in assessing the symptom criteria for depression were clearly superior to the three studies reviewed above that used CTT methodology in that the unidimensionality for the symptom criteria was checked, either indirectly through extensive analysis of the fit of different kinds of IRT models that were used (Reiser, 1989) or directly through the use of factor analytic methodology (Aggen et al., 2005). The authors of both IRT studies began with explicit assumption that the symptom criteria for depression may be driven by an underlying continuous dimension and they used IRT to determine the precise relationship between item response and the underlying latent construct of depression. However, each of the two IRT studies did have some limitations centered mainly on the representativeness of the data sets that were used. The Reiser (1989) study was based on data from the ECA and mainly used data from only one metropolitan city. The Aggen et al. (2005) study had a biased data set in that all subjects were female Caucasian twins from Virginia.

Conclusions

It appears that the use of IRT models for investigating psychopathological constructs and the diagnosis of psychopathology holds great promise. As has been demonstrated above, IRT models can provide a potentially rich and innovative perspective on noncognitive measures and constructs. It is unfortunate that the IRT framework has not been used more extensively in psychodiagnostic research.

The current proposed study will add to the literature on the psychometric modeling of the psychodiagnostic symptom criteria of depression. It will use CTT and IRT techniques on data from a nationally representative epidemiological data set in order to elaborate the relationship between each individual symptom of depression and the

underlying construct of depression. As part of the psychometric analysis, the proposed study will assess the dimensionality of the set of symptom criteria for depression and it will use the results of the IRT analysis to critically examine the cutpoint criteria of five symptoms. Ultimately, the proposed study will give a good indication of the overall quality of the set of symptom criteria for depression through the perspective of the CTT and IRT psychometric paradigms.

CHAPTER III: STUDY DESIGN AND METHODOLOGY

NCS-R Data Set

The proposed study will utilize data collected by the National Comorbidity Study – Replication (NCS-R) (Kessler & Merikangas, 2004). The NCS-R was an epidemiological study of the prevalence and severity of major psychiatric disorders as defined by the diagnostic criteria of DSM-IV (Kessler et al., 2004; Kessler & Merikangas, 2004). The NCS-R was carried out in a timeframe between February 2001 and April 2003 (Kessler et al., 2004). The predecessor to the NCS-R was the original NCS, which used the diagnostic criteria of DSM-IIIR (Kessler & Merikangas, 2004).

Participants

> The NCS-R
>
> was designed to be representative of English-speaking adults ages 18 or older living in the non-institutionalized civilian household population of the coterminous US (excluding Alaska and Hawaii). (Kessler et al., 2004, p. 72)

Design and Procedures

The NCS-R utilized a complex design survey methodology in a multistage probability sampling framework (Kalton, 1983; Kessler et al., 2004). Participants were selected in a four stage process. The first stage involved creating a set of "primary sampling units (PSUs)" (Kessler et al., 2004, p. 74) across the map of the 48 contiguous United States, each of which consisted "of all counties in a census-defined metropolitan statistical area (MSA) or, in the case of counties not in an MSA, of individual counties" (Kessler et al., 2004, p. 74). The PSUs that were created for the NCS-R were judged by the design team to be "representative of the population" (Kessler et al., 2004, p. 74). A

total of 84 PSUs were selected for the administration of the NCS-R (Kessler et al., 2004). The second stage involved creating geographical groupings of approximately 50 to 100 "housing units (HUs)" (Kessler et al., 2004, p. 74) within each PSU based on U.S. Census 2000 data and then choosing approximately 12 of the geographical groupings from each PSU, for a grand total of 1,001 geographical groupings of HUs (Kessler et al., 2004). The third stage involved investigating the addresses of all HUs in each selected geographical grouping for the purpose of updating address records from the U.S. Census and/or the previous NCS survey (Kessler et al., 2004).

> In order to adjust for discrepancies between expected and observed numbers of HUs, a random sample of HUs was selected that equals 10 * O/E, where O is the observed number of households listed in the segment and E is the number of HUs expected in the segment from the Census data files. (Kessler et al., 2004, p. 74)

The fourth stage involved randomly selecting one or two adult English speaking members of each HU to participate in the NCS-R (Kessler et al., 2004).

For subsequent analyses, the 84 PSUs were paired together based on matching criteria into 42 strata (Kessler et al., 2004). Thus, each stratum was initially constructed out of two PSUs (Kessler et al., 2004). The individuals assigned to each stratum were then randomly split into two groups (Kessler et al., 2004; National Comorbidity Survey, n.d.) that were subsequently treated as "sampling error calculation units (SECUs)" (Kessler et al., 2004, p. 86).

Instrument

The NCS-R used a version of the Composite International Diagnostic Interview (CIDI; Kessler et al., 2004; Kessler & Üstün, 2004) that was modified by the NCS-R

design team. The CIDI was initially developed by the WHO to be used as a structured interview survey instrument in epidemiological studies of the prevalence rates of psychiatric disorders in different countries (Kessler & Üstün, 2004). The CIDI was designed to be administered in a face-to-face setting by trained professional interviewers who were not mental health professionals (Kessler & Üstün, 2004). The format of the CIDI consisted of questions that tapped into various psychiatric disorders using "skip logic" (Kessler et al., 2004, p. 70). The skip logic technique for survey instruments consists of presenting certain exploratory questions that initially probe for the possibility of particular psychiatric disorders (Kessler et al., 2004; Kessler & Üstün, 2004). If a participant answers negatively to those probe questions for a particular psychiatric disorder, then they are "skipped out" of the rest of the questions for that disorder (Kessler et al., 2004; Kessler & Üstün, 2004). One of the benefits of using skip logic is that it decreases participant fatigue by reducing the number of questions that participants have to answer (Kessler et al., 2004; Kessler & Üstün, 2004).

The NCS-R design team modified the original CIDI in order to improve data collection and to increase the reliability and validity of the instrument (Kessler & Üstün, 2004). Especially noteworthy modifications were a set of changes to the format of the CIDI that addressed the issue of participant fatigue over the course of the administration of the instrument (Kessler & Üstün, 2004). Participant fatigue was one of the most salient problems in administering the original version of the instrument that affected data quality (Kessler et al., 2004; Kessler & Üstün, 2004). At a minimum, the time to completion for the CIDI by participants with no diagnosable psychiatric disorders is 90 minutes (Kessler et al., 2004). The time to completion for the CIDI increases as a direct

function of the number and/or severity of lifetime diagnosable psychiatric disorders reported by participants (Kessler et al., 2004). For participants with severe cases and/or numerous different types of lifetime psychiatric disorders, the time to completion for the CIDI can be as long as five or six hours (Kessler et al., 2004). Data quality can also suffer because many participants catch on to the skip logic of the CIDI and start responding negatively to the probe questions that are placed throughout the CIDI (Kessler et al., 2004; Kessler & Üstün, 2004). In order to avoid the problem of participant fatigue causing negative responses to probe questions, the NCS-R design team decided to move all the probe questions to the initial screening section of the interview (NCS-R Section 2: Screener; Kessler, n.d. a; Kessler et al., 2004; Kessler & Üstün, 2004).

Another innovation introduced by the NCS-R design team was to split the survey into two parts (Part I and Part II) (Kessler et al., 2004; Kessler & Merikangas, 2004; Kessler & Üstün, 2004). Part I was administered to all participants and it covered basic psychiatric disorders such as anxiety, depression and mania (Kessler & Merikangas, 2004; Kessler & Üstün, 2004). Part II was administered chiefly to participants who had in Part I reported suffering some level of psychopathology and/or had received psychiatric treatment during the course of their lives (Kessler et al., 2004; Kessler & Üstün, 2004). Part II was also administered to a select group of randomly chosen participants that never reported any sort of psychopathology and/or treatment (Kessler et al., 2004; Kessler & Üstün, 2004). Part II delved more deeply into the participant's sociodemographics, risk factors, life history of coping with disorders, and the assessment of additional disorders that were not assessed in Part I (Kessler et al., 2004; Kessler &

Üstün, 2004). The total number of participants in Part I was 9,282 and the total number of participants in Part II was 5,692 (Kessler et al., 2004).

The proposed study will use the data collected by the NCS-R found in the section of the CIDI that assessed depression (NCS-R Section 3: Depression; Kessler, n.d. b). The initial screening section of the NCS-R (NCS-R Section 2: Screener; Kessler, n.d. a) asked three probe questions:

> *SC21. Have you ever in your life had a period lasting several days or longer when most of the day you felt sad, empty or depressed?
>
> *SC22. Have you ever had a period lasting several days or longer when most of the day you were very discouraged about how things were going in your life?
>
> *SC23. Have you ever had a period lasting several days or longer when you lost interest in most things you usually enjoy like work, hobbies, and personal relationships? (p. 30)

If a participant answered affirmatively to any of the above three questions, they were then flagged as someone who would be given the depression section. The depression section had an initial series of further probe questions concerning the duration and severity of a depressive episode (Kessler, n.d. b). If a participant answered affirmatively to any of those questions, then they were given the complete battery of questions that assessed depression. If a participant answered negatively to any of the probe questions at the beginning of the depression section, then they were skipped out of the rest of the section. The responses to the remainder of the questions in the depression section were designed to create a symptom profile of a major depressive episode as defined by DSM-IV (Kessler, n.d. c).

Analyses

Because of the use of skip outs by the NCS-R design team in order to minimize participant fatigue, 1,978 participants out of the total of 9,282 participants, or 21.3%, were given the CIDI section on depression (Kessler et al., 2004). Based on the content of the initial probe questions found in the screening and depression sections (Kessler, n.d. a, n.d. b, n.d. c), it can be inferred that the participants for which there is data on symptoms of depression are individuals who had some level of distress related to low mood at some point in their lives and therefore they are individuals who most likely would have been or currently are candidates for a diagnosis of depression. Thus, the 1,978 individuals who took the CIDI section on depression most likely resemble the type of individuals who would be found in clinical trials of depression or in an outpatient treatment facility, for instance. Also, another issue to consider is that by definition, the DSM-IV criteria for a MDE do not consider anyone a candidate for a diagnosis of depression if they do not meet one of the two gate symptom criteria of depressed mood and/or anhedonia (APA, 2000). An individual can have all seven of nine nongate symptoms of depression but if they do not have at least one of the initial gate criteria, then they cannot be diagnosed with depression.

The nine symptoms of a MDE will be constructed from the responses to the questions found in the section on depression in the CIDI. The algorithms for constructing the symptom criteria of depression are found on the website of the National Comorbidity Survey website (http://www.hcp.med.harvard.edu/ncs/index.php).

Classical Test Theory

The initial step for the CTT (and also IRT) analysis of the symptoms of a MDE will be to determine the unidimensionality of the nine symptom criteria using exploratory factor analysis on a matrix of tetrachoric correlations of the nine symptoms. Most, if not all, tests cannot be perfectly unidimensional, however, Reckase (1979) has shown that as long as a test has one strong dominant factor, an IRT analysis will most likely work off that one dominant factor.

An examination of the factor structure will reveal the dimensionality of the nine symptoms. If there is one strong dominant factor as revealed by the eigenvalues and the scree plot, then all nine symptoms will be concurrently analyzed as one instrument. If two or more strong factors are revealed, then the subsequent analyses will treat the sets of symptom criteria associated with each factor separately. The exploratory factor analysis will be conducted using Mplus, which has the capability to compute a tetrachoric correlation matrix for the factor analysis (Muthen & Muthen, 1998-2006).

The difficulty value for each symptom will be calculated (Crocker & Algina, 1986). Cronbach's α will be used to determine the overall reliability of the set of symptom criteria (Crocker & Algina, 1986). Three types of correlations (Pearson's, Spearman's Rho and Kendall's tau-b) between each symptom and the total number of symptoms will be calculated (Crocker & Algina, 1986).

IRT

Two sets of IRT analyses will be used to assess the symptom criteria of depression in DSM-IV. The first set will consist of the 1PL, 2PL, and 3PL IRT models. BILOG (du Toit, 2003) will be used to fit the three IRT models with all default settings in place. The

fit of each model will be checked by using the likelihood ratio test and the Akaikie Information Criteria (AIC) (de Ayala, 2009). The likelihood ratio test involves testing the difference between two different log likelihood values for two different statistical models, one of which is nested within the other (de Ayala, 2009). The difference between both log likelihood values is tested against a chi-square distribution, with degrees of freedom equal to the difference between the number of parameters in both models (de Ayala, 2009). The 1PL model is nested within a 2PL model, and the 1PL and 2PL models are nested within the 3PL model. The likelihood ratio test will be carried out between all nested IRT models. The AIC is a test of fit that takes into account the number of parameters that is estimated for each model (Rost, 1997). The AIC test for each model will be computed using the formula:

$$\mathrm{AIC} = -2 \log L + 2k, \tag{12}$$

where log(L) is the log likelihood value for a particular model, and k is the number of estimated parameters for each IRT model (Rost, 1997). For the 1PL model, the number of parameters is 10, i.e., nine *b* parameters and one *a* parameter that has been constrained to equivalence across all indicator items (de Ayala, 2009). For the 2PL model, the number of parameters is 18, i.e., nine *a* parameters and nine *b* parameters. For the 3PL model, the number of parameters is 27, i.e., nine *a* parameters, nine *b* parameters, and nine *c* parameters. Graphical plots of the item information curves, test information functions, test standard error of estimate, and the total test characteristic curve will be created (de Ayala, 2009; du Toit, 2003).

The marginal posterior probability value that is produced during the scoring process for each individual in an IRT model can be used as an index of fit for each individual's score (du Toit, 2003; de Ayala, 2009). The marginal posterior probability value is reflective of the consistency of the score pattern of each individual (de Ayala, 2009). Higher values of the marginal posterior probability value indicates an answer profile that has a higher degree of consistency within the sample. The log(marginal posterior probability) value for each individual will be plotted in a jittered scatterplot against the total number of symptoms.

The response data for the nine MDE symptoms from the lower 25% of the log(marginal posterior probability) values in the sample will be subjected to a two stage cluster analysis in order to determine whether there are any patterns of symptom responding that are associated with lower marginal posterior probability values. The first stage will involve carrying out an average linkage cluster analysis (Everitt, 1980) using SAS PROC CLUSTER, the results of which will be plotted in a standard dendrogram diagram. Visual inspection of the dendrogram will be used to determine which clusters are robust enough to be retained in further analysis. The centroids of the clusters retained from the visual inspection will be then used as seeds in a k-means analysis (Hastie, Tibshirani, & Friedman, 2001), which will be carried out using SAS PROC FASTCLUS. A profile for each cluster from the k-means analysis that consists of the means of each symptom response will be constructed and plotted.

A set of initial analyses showed that the reversed keyed 3PL could not converge on a solution unless the depressed mood criterion was removed. This was most likely due to the high rate of endorsement of depressed mood, which was no doubt the result of

depressed mood being one of the gate symptoms for a diagnosis of depression. It was therefore determined that in order to determine the fit of the reversed keyed 3PL model as compared to the nonreversed keyed models using the likelihood ratio tests and the AIC, a second set of 1PL, 2PL, and 3PL IRT models would be run on the symptom criteria without the depressed mood criterion.

CHAPTER IV: STUDY RESULTS

Classical test theory analysis

Factor analysis

In order to ascertain the underlying dimensional structure of the nine symptom criteria for a major depressive episode, a factor analysis was run based on tetrachoric correlations using Mplus (Muthen & Muthen, 1998-2006). The screeplot showed one dominant factor, and therefore, a one factor solution of the factor analysis was retained. All nine symptom criteria will be treated as one instrument for purposes of the psychometric analyses. The estimated factor loadings for the one factor solution are found in Table 1. The loadings reveal that all of the symptoms have a moderate relationship to the underlying factor, with depressed mood having the weakest loading (.394) and thinking/concentration problems having the strongest loading (.699). Cronbach's alpha for the entire set of nine symptom criteria was .583, which indicates only a moderate level of reliability for the total of the nine MDE symptom criteria.

Table 1
Estimated Factor Loadings of the Nine DSM-IV Symptom Criteria for a Major Depressive Episode Based on Tetrachoric Correlations

Criteria	Loadings
1. Depressed Mood	0.394
2. Anhedonia	0.545
3. Weight/Appetite Problems	0.471
4. Sleep Problems	0.596
5. Psychomotor Problems	0.572
6. Fatigue	0.555
7. Worthlessness/Guilt	0.665
8. Thinking/Concentration Problems	0.699
9. Suicidal Tendencies	0.441

Item difficulty

The distribution of the total summed score of the nine symptom criteria is shown in Figure 1. The difficulty, or *p*-value, for each symptom criterion was computed (Table 2).

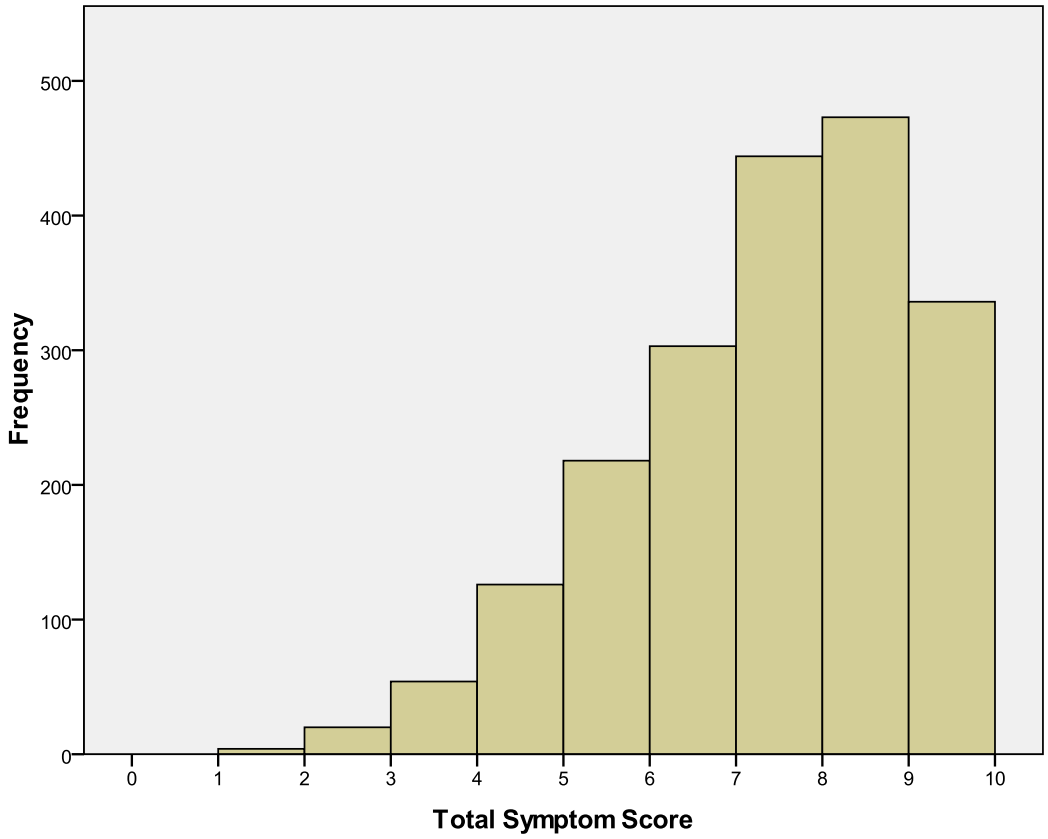

Figure 1. Distribution of the total symptom score of the nine symptom criteria for a MDE.

Table 2
Difficulty Levels for the Nine DSM-IV Symptom Criteria for a Major Depressive Episode

Criteria	Difficulty
1. Depressed Mood*	0.987
2. Anhedonia*	0.835
3. Weight/Appetite Problems	0.835
4. Sleep Problems	0.901
5. Psychomotor Problems	0.481
6. Fatigue	0.825
7. Worthlessness/Guilt	0.415
8. Thinking/Concentration Problems	0.859
9. Suicidal Tendencies	0.677

* denotes gate criterion

The analysis of the difficulty levels of the symptoms revealed that the most endorsed item is depressed mood (p = .987). The next highest endorsed symptom was sleep problems (p = .901). Four symptoms (thinking/concentration problems, anhedonia, weight/appetite problems, and fatigue) all had endorsement rates in the low to mid .80's. The remaining three symptoms (suicidal tendencies, psychomotor problems, and worthlessness/guilt) all had moderate levels of endorsement in the .60's and .40's. Also of note is that six out of the nine symptoms (depressed mood, anhedonia, weight/appetite problems, sleep problems, fatigue, and thinking/concentration problems) were endorsed by more than 80% of the sample.

Item-Total Score Correlations

The total summed score of the nine MDE symptoms was correlated with each individual item using parametric (Pearson's) and nonparametric (Spearman's Rho and Kendall's tau-b) correlations (Table 3).

Table 3
Correlations of Individual DSM-IV Symptom Criteria for a Major Depressive Episode with the Total Summed Score of All Symptom Criteria

MDE Symptom Criteria	Pearson's Correlation	Spearman's Rho	Kendall's tau-b
1. Depressed Mood	.150	.135	.119
2. Anhedonia	.529	.484	.428
3. Weight/Appetite Problems	.450	.415	.367
4. Sleep Problems	.442	.392	.347
5. Psychomotor Problems	.572	.598	.529
6. Fatigue	.470	.428	.379
7. Worthlessness/Guilt	.601	.645	.570
8. Thinking/Concentration Problems	.523	.460	.406
9. Suicidal Tendencies	.501	.507	.448

The item-total score correlations appear to cluster into roughly three groups. The first group consists of only depressed mood, which has an extremely low item-total correlation. The reason for the low item-total correlation for depressed mood is most likely due to the extremely high level of endorsement for this symptom in the NCS sample. In the CTT paradigm, a low item-total correlation is expected for an indicator with an extreme level of difficulty (Crocker & Algina, 1986). The second group of item-total score correlations consists of sleep problems, weight/appetite problems, and fatigue, and these symptoms had moderate item-total correlations in the .40's. Two of these symptoms, sleep and weight/appetite problems, are the two vegetative symptoms in the total set of nine symptoms. Fatigue is a physical symptom. The third group of item-total score correlations consisted of suicidal tendencies, thinking/concentration problems, anhedonia, psychomotor problems, and worthlessness/guilt. These symptoms had higher item-total score correlations in the range between .501 and .601, and they are emotional and cognitive in nature. Overall, the vegetative and physical symptoms have lower item-total correlations than the emotional and cognitive symptoms, excepting depressed mood.

Item response theory analysis

1PL, 2PL, 3PL, & reversed key 3PL IRT models

A set of IRT analyses consisting of the 1PL, 2PL, 3PL and reversed key 3PL models was run. Table 4 contains the log likelihood and AIC values for the 1PL, 2PL, and 3PL IRT models as well as a series of chi-square tests of fit (1PL vs. 2PL and 3PL; 2PL vs. 3PL) using the log likelihood values. Table 5 contains the log likelihood and AIC values for the reversed key 3PL IRT model as well as a series of chi-square tests of fit (1PL vs. 2PL, 3PL, and reversed key 3PL; 2PL vs. 3PL and reversed key 3PL) using the log likelihood values. Because the reversed key 3PL model could only be fit without the depressed mood criteria, a series of 1PL, 2PL, and 3PL IRT models were created without the depressed mood criteria included for use in the log likelihood tests of fit with the reversed key 3PL IRT model.

The results of the fit tests show that for the nonreversed keyed criteria the 2PL model fits best and therefore it will be considered as the primary IRT model in the results and discussion. However, the 3PL model will also be considered for several reasons. First, it is of a priori interest to examine the *c* parameters of the 3PL model in order to estimate the level of responding for individual diagnostic criteria at low levels of depression. A substantial *c* parameter value for an individual symptom indicates that there is a bias to detect that symptom in the lower range of depression at probability levels greater than expected. Second, the AIC for the 3PL model was second ranked among the three IRT models examined and the log likelihood test determined that it was significantly better fitting than the 1PL model. Third, the log likelihood test between the 3PL and 2PL models was not statistically significant and the 3PL model significantly fit

better than the 1PL model. Finally, the 3PL model among all the four reversed key IRT models clearly fit the best, and it is useful to compare the nonreversed and reversed keyed 3PL models with each other.

Table 4
Fit Statistics for the 1PL, 2PL, and 3PL IRT Models for the Major Depressive Episode

IRT Model	Number of Parameters	-2lnL value	χ^2 Difference with 1PL	χ^2 Difference with 2PL	AIC
1PL	10	15433.479			15453.48
2PL	18	15380.048	-53.4317*		15416.05
3PL	27	15390.209	-43.2705*	10.1612	15444.21

* $p < .05$.

Table 5
Fit Statistics for the Reversed Keyed 3-PL IRT Model and the Associated 1PL, 2PL, and 3PL IRT Models for the Major Depressive Episode

IRT Model	Number of Parameters	-2lnL value	χ^2 Difference with 1PL	χ^2 Difference with 2PL	AIC
1PL	9	15301.390			15319.39
2PL	17	15266.793	-34.597*		15300.79
3PL	26	15602.957	301.567*	336.164*	15654.96
rev 3PL	26	15139.229	-162.1608*	-127.5638*	15191.23

Note. All models were run without the depressed mood criterion.
* $p < .05$.

The item statistics for the 2PL, 3PL, and reversed keyed 3PL models are shown in Tables 6-8. The graphical representation of the item characteristic curves, item information curves, test information curves, test standard error curves, and total test characteristic curves for the 2PL model is shown in Figures 2-5. The item characteristic curves for the 3PL and the reversed keyed 3PL models are shown in Figures 6 and 7,

respectively. More detailed graphical results of the 1PL, 3PL and the reversed keyed 3PL models are shown in the Appendix.

Table 6
2PL IRT Parameters for the Major Depressive Episode Symptom Criteria

Symptom Criteria	parameters			
	a	*a* SE	*b*	*b* SE
1. Depressed Mood	0.628	0.177	-4.679	1.084
2. Anhedonia	0.869	0.084	-1.532	0.101
3. Weight/Appetite Problems	0.576	0.059	-2.02	0.169
4. Sleep Problems	0.766	0.086	-2.139	0.165
5. Psychomotor Problems	0.642	0.055	0.046	0.052
6. Fatigue	0.63	0.058	-1.75	0.131
7. Worthlessness/Guilt	0.798	0.068	0.336	0.049
8. Thinking / Concetration Problems	1.003	0.099	-1.613	0.096
9. Suicidal Tendencies	0.475	0.046	-1.042	0.104

Table 7
3PL IRT Parameters for the Major Depressive Episode Symptom Criteria

Symptom Criteria	parameters					
	a	*a* SE	*b*	*b* SE	*c*	*c* SE
1. Depressed Mood	0.828	0.212	-3.816	0.741	0.5	0.112
2. Anhedonia	1.483	0.346	-0.616	0.194	0.468	0.077
3. Weight/Appetite Problems	0.805	0.141	-0.84	0.322	0.47	0.095
4. Sleep Problems	1.15	0.214	-1.128	0.264	0.5	0.098
5. Psychomotor Problems	0.805	0.12	0.402	0.109	0.153	0.041
6. Fatigue	0.754	0.102	-0.986	0.272	0.365	0.092
7. Worthlessness/Guilt	1.217	0.253	0.604	0.071	0.144	0.031
8. Thinking / Concetration Problems	1.399	0.261	-0.976	0.196	0.397	0.092
9. Suicidal Tendencies	1.055	0.274	0.335	0.164	0.457	0.049

Table 8
Reversed Key 3PL IRT Parameters for the Major Depressive Episode Symptom Criteria

Symptom Criteria	parameters					
	a	*a* SE	*b*	*b* SE	*c*	*c* SE
1. Depressed Mood	-	-	-	-	-	-
2. Anhedonia	0.859	0.165	1.517	0.097	0	0.027
3. Weight/Appetite Problems	0.569	0.140	2.01	0.172	0	0.042
4. Sleep Problems	0.799	0.166	2.058	0.170	0	0.018
5. Psychomotor Problems	0.804	0.176	0.293	0.256	0.14	0.104
6. Fatigue	0.893	0.189	1.714	0.120	0.05	0.022
7. Worthlessness/Guilt	0.783	0.186	-0.342	0.488	0	0.255
8. Thinking / Concetration Problems	0.988	0.188	1.612	0.100	0	0.018
9. Suicidal Tendencies	0.443	0.118	1.094	0.370	0	0.125

Note: The reversed key 3PL IRT model could only be fit if the depressed mood criterion was removed.

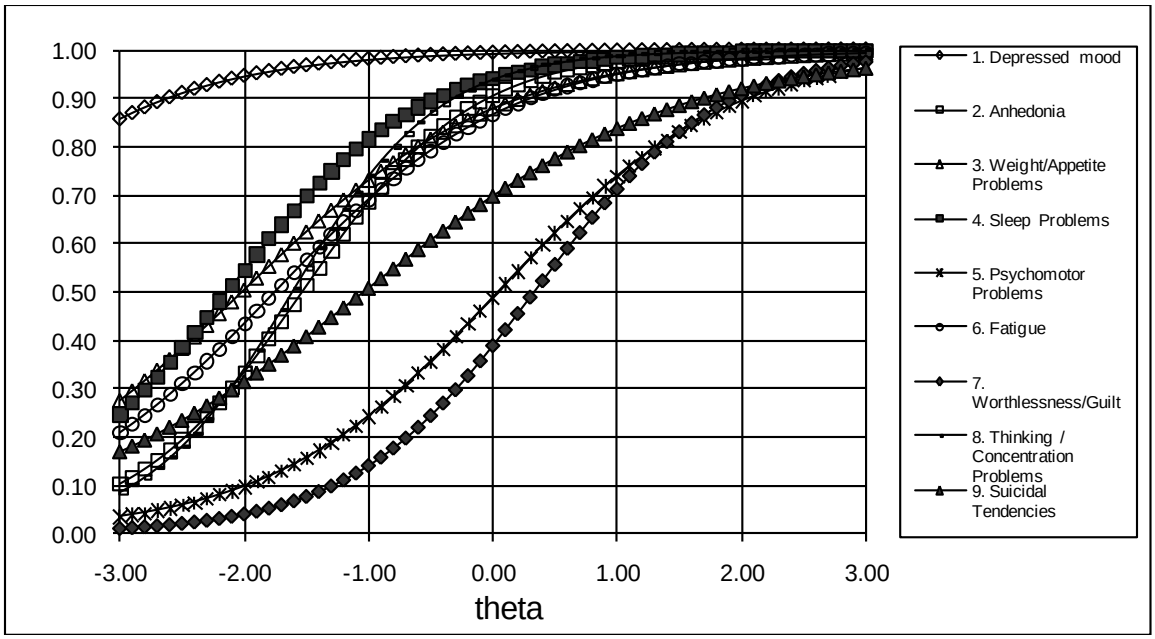

Figure 2. Item characteristic curves for the 2PL IRT model.

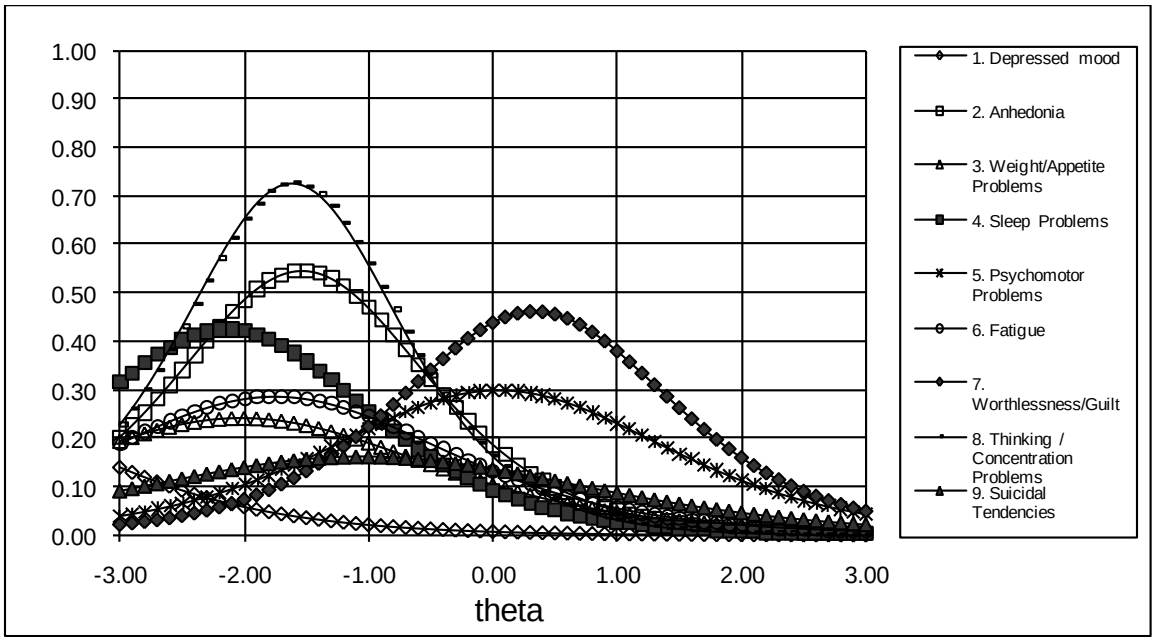

Figure 3. Item information curves for 2PL IRT model.

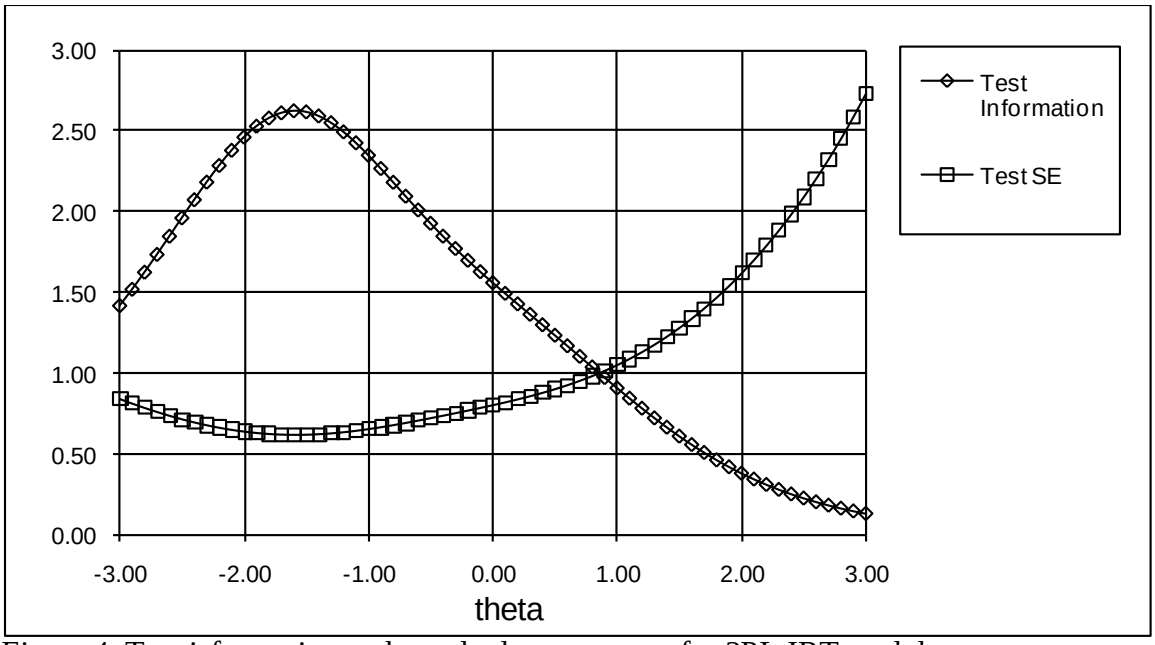

Figure 4. Test information and standard error curves for 2PL IRT model.

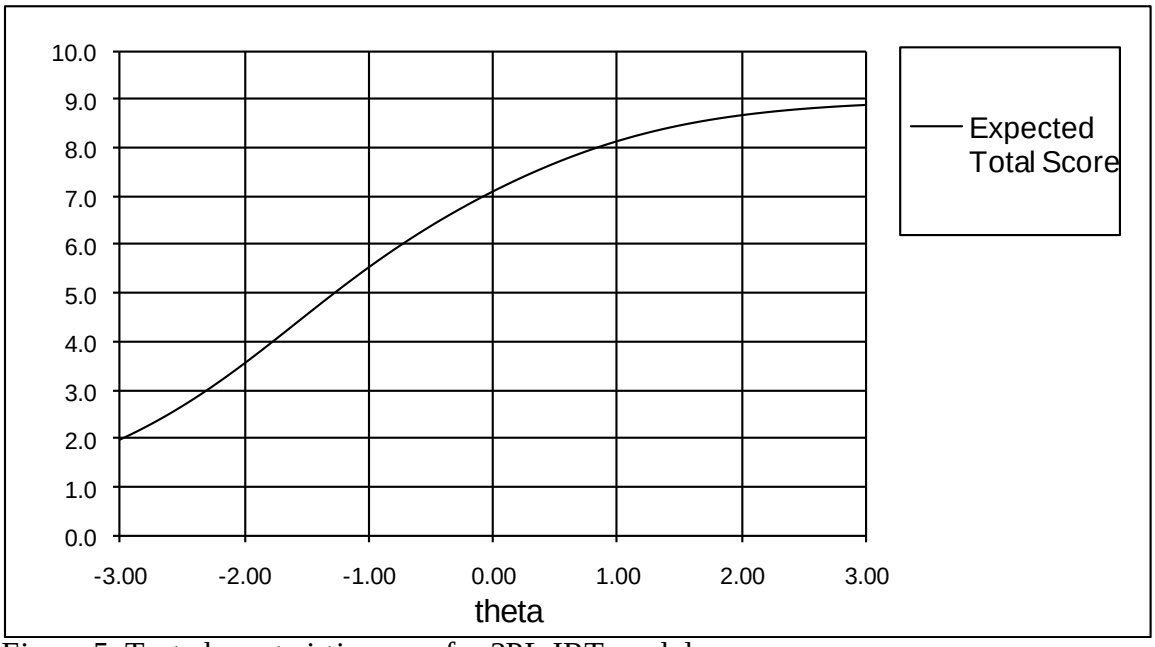

Figure 5. Test characteristic curve for 2PL IRT model.

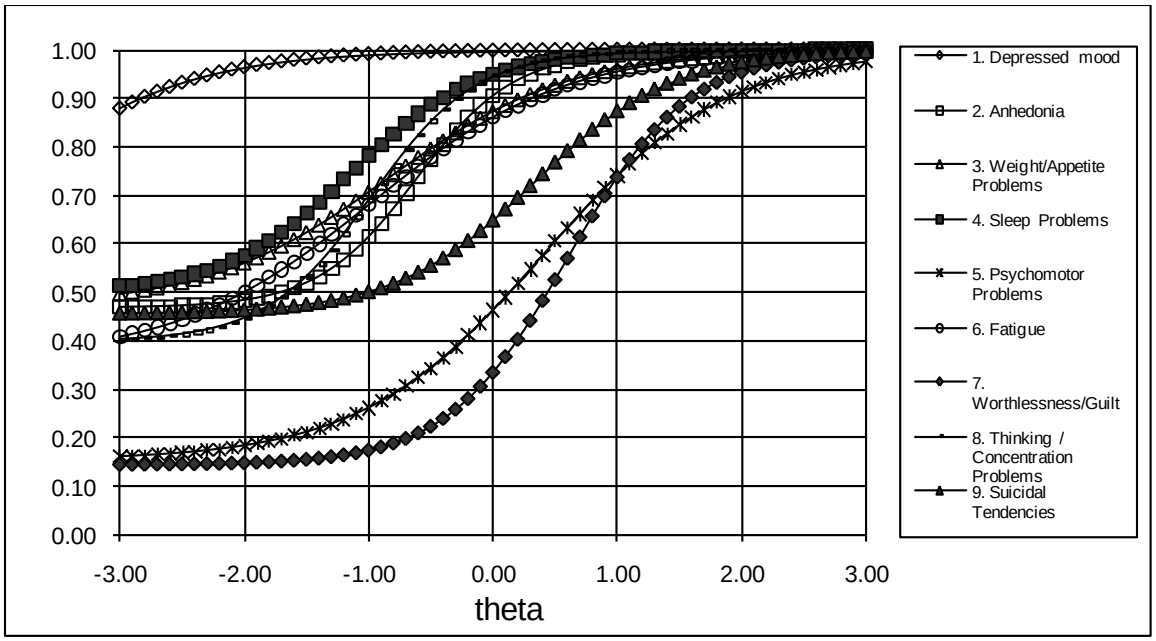

Figure 6. Item characteristic curves for the 3PL IRT model.

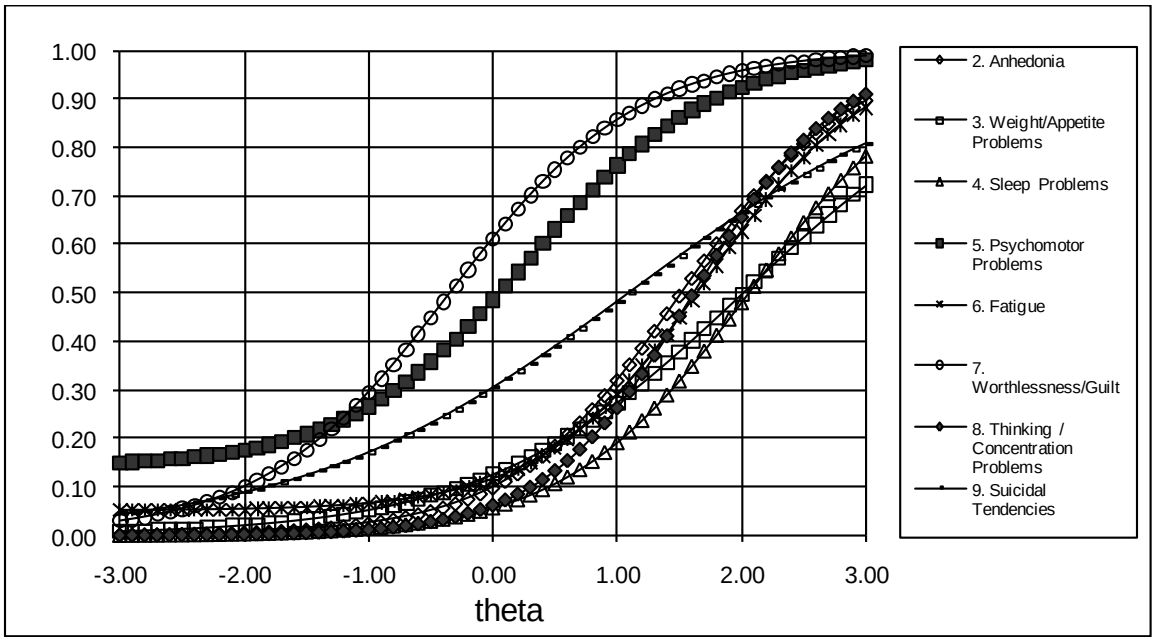

Figure 7. Item characteristic curves for the reversed key 3PL IRT model.

One of the most immediately striking results that emerged from the 2PL IRT model is that the ICC for the criterion of depressed mood, which is one of the two gate criteria, had an extremely low difficulty parameter (Table 6 and Figure 2). Anhedonia, which is the other gate criterion, had a much higher difficulty level (Table 6 and Figure 2). Another striking result of the 2PL IRT model was that five of the nine symptom criteria (sleep problems, thinking/concentration problems, weight/appetite problems, anhedonia, and fatigue) had remarkably similar ICC profiles that appeared to cluster together in the plot of the ICCs (Figure 2). The overall difficulty level of this cluster of five symptom criteria was moderately low. After this cluster of the five symptom criteria, the next most difficult diagnostic criterion was suicidal tendencies and it had the lowest discriminating parameter among the nine symptoms (Table 6 and Figure 2). The most difficult diagnostic symptoms were psychomotor problems and worthlessness/guilt (Table 6). Their ICCs clustered together toward the upper end of the latent spectrum of depression (Figure 2).

For the 2PL model, the item information curves (Figure 3) show that thinking/concentration problems and anhedonia are associated with the largest amount of information for the construct of depression. The information curves for both symptoms peak at a theta value of approximately -1.5. The information curves for sleep problems, worthlessness/guilt and psychomotor problems are also fairly substantial, as compared to the information curves for the other symptoms. Taken as a whole, the information curves for all symptoms show that most of the symptoms contribute the majority of their information before a theta level of approximately -0.5. The exception to this pattern is the information curves for worthlessness/guilt and psychomotor problems, the curves for

which peak at theta levels at approximately .5 and 0, respectively. The peak of the information curve for depressed mood is far below the -3.0 theta value due to its extremely low difficulty level. The ICC for depressed mood indicates that this symptom actually contributes little information to the overall diagnosis of depression once an individual meets the minimal criteria of depressed mood (Figure 2). Also of note is that the information curve for suicidal tendencies has the lowest maximum value of all the nine symptom criteria, which indicates that, within the theta range of -3 to 3, it contributes least to the diagnosis of depression. The test information curve (Figure 4) for the 2PL model shows that the nine symptom criteria provide the maximum information about an individual's location on the depressive spectrum at a theta value of approximately -1.5.

For the 3PL IRT model, the main interest was in the *c* parameters in order to determine the propensity for clinical symptoms to be found in the lower range of depression at levels greater than expected. The 3PL model found that the *c* parameters of the 3PL model are fairly substantial for the majority of items. Six out of the nine symptoms (depressed mood, anhedonia, weight/appetite problems, sleep problems, thinking/concentration problems, and suicidal tendencies) have *c* parameters that are in between the range of .4 and .5. Fatigue has a *c* parameter of .37. The two most difficult symptoms, psychomotor problems and worthlessness/guilt, have the two lowest *c* parameters in the model, .15 and .14, respectively. The difficulty and discrimination parameters of the ICCs of the 3PL model had patterns similar to the 2PL model.

For the reversed keyed 3PL model, the main interest was also in the *c* parameters in order to determine the propensity for clinical symptoms to be found in the higher range of depression at levels less than expected. The reversed keyed 3PL model had zero *c* parameters for all but two symptoms. The reversed keyed symptoms of psychomotor problems and fatigue had *c* parameters of .14 and .05, respectively, i.e., those two symptoms had small nonunity asymptotes in the nonreversed keyed direction.

For the reversed keyed 3PL model, only models without the depressed mood criterion converged on a solution, so the 1PL, 2PL, and 3PL models were rerun without the depressed mood criterion to be used for comparison purposes with the reversed key 3PL model. Among the models without the depressed mood criterion, the reversed keyed 3PL model fit the best, according to both the likelihood ratio tests and the AIC.

The parameter values for the complete IRT models and the IRT models without the depressed mood criteria were compared in order to determine the degree to which the symptom parameters can be compared across the two different kinds of models. For the 1PL IRT, the average difference for all the *b* parameter values of the eight symptom criteria included in the reversed key 3PL IRT model with the *b* parameter values in the complete IRT model (with the average difference in the standard errors in parentheses) was 0.0011 (0.1243). For the 2PL IRT, the average difference for the nondepressed mood symptom criteria between both kinds of models (with the average difference in the standard errors in parentheses) for the *a* and *b* parameter values was 0.0005 (0.1389) and 0.0017 (0.2160), respectively. For the 3PL IRT, the average difference for the nondepressed mood symptom criteria between both kinds of models (with the average difference in the standard errors in parentheses) for the *a*, *b*, and *c* parameter values was -

0.0092 (0.4245), -0.0115 (0.3999), and -0.0026 (0.1446) respectively. Thus, the *a*, *b*, and *c* parameters for the IRT models without depressed mood were almost identical to their respective parameter values from the nine symptom version of the IRT model.

Marginal posterior probability value

The natural log of the marginal posterior probability value for the 2PL IRT model was plotted against the total number of MDE symptom criteria endorsed (Figure 8).

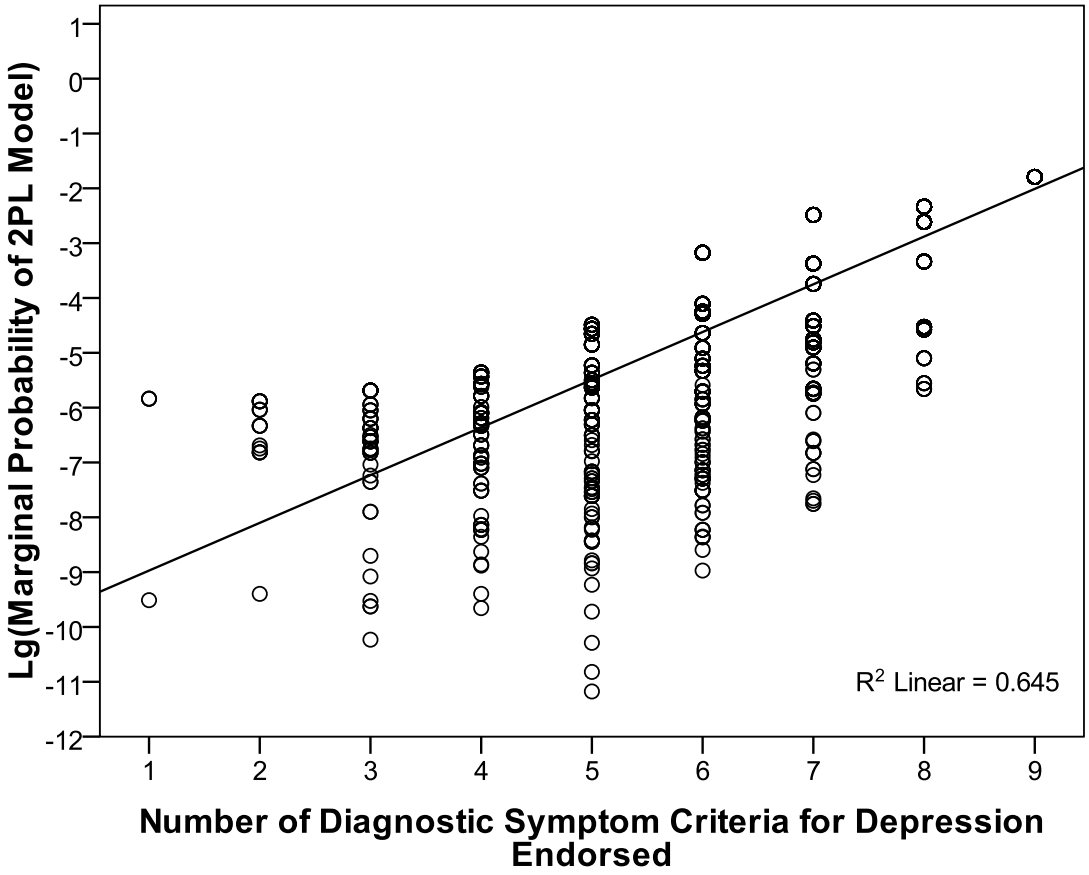

Figure 8. Scatterplot of the natural log of the marginal posterior probability value of the 2PL IRT model regressed against the total number of major depressive episode symptom criteria endorsed.

The scatterplot in Figure 8 shows that there was a strong relationship between the 2PL marginal posterior probability value and the total number of MDE symptom criteria endorsed. However, the scatterplot also reveals that there was a wide range of marginal posterior probability values across individuals, especially those that endorsed between three to eight MDE symptom criteria, inclusive. These results indicate that the consistency of the responses in general increases with the total number of symptom criteria endorsed. However, even for individuals with a total of seven or eight symptoms, there is still a notable degree of scatter of the marginal posterior probability values, which indicates that there were a large amount of individuals who had inconsistent patterns of responding even with a high number of total symptoms endorsed. The symptom patterns of individuals who had inconsistent patterns of responding was assessed with a cluster analysis.

Cluster Analysis

A cluster analysis using average linkage was conducted on the responses of the individuals for the lowest 25% of the marginal posterior probability value distribution. The dendrogram of the final solution is shown in Figure 9. Clusters to be retained were chosen through visual inspection of the dendrogram, with a total of eight clusters retained (arrows in Figure 9).

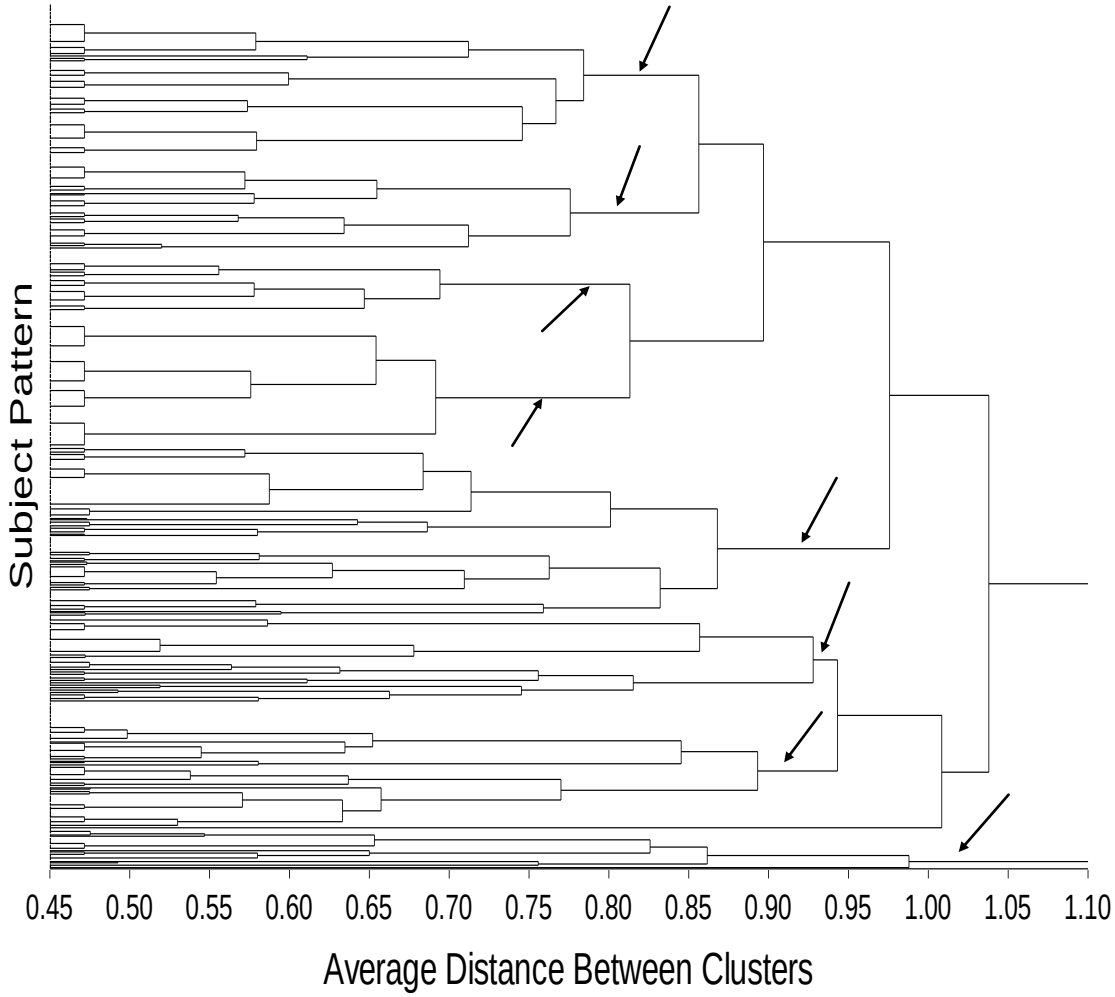

Figure 9. Dendrogram of an average linkage cluster analysis of the nine major depressive episode symptom criteria for individuals who were in the lowest 25% of the marginal posterior probability value distribution.

The means of the MDE symptom criteria for each of the eight clusters were used to initially seed a k-means cluster analysis. The number of individuals in Clusters 1-8 was: 50, 61, 87, 20, 79, 65, 43, and 87, respectively. The result of the k-means analysis is shown in Table 9 and Figures 10 and 11.

Table 9
Statistics of the k-Means Analysis for the Nine DSM-IV Major Depressive Episode Symptom Criteria

Symptom Criteria	Total STD	Within STD	R^2	$R^2/(1- R^2)$
1	0.21	0.08	0.87	6.45
2	0.50	0.45	0.20	0.25
3	0.50	0.34	0.52	1.10
4	0.48	0.29	0.64	1.79
5	0.45	0.28	0.62	1.66
6	0.50	0.44	0.25	0.33
7	0.43	0.14	0.89	8.50
8	0.50	0.45	0.19	0.23
9	0.50	0.46	0.18	0.22
OVER-ALL	0.46	0.35	0.43	0.74

Pseudo F Statistic = 51.36
Approximate Expected Over-All R^2 = 0.41112
Cubic Clustering Criterion = 3.116

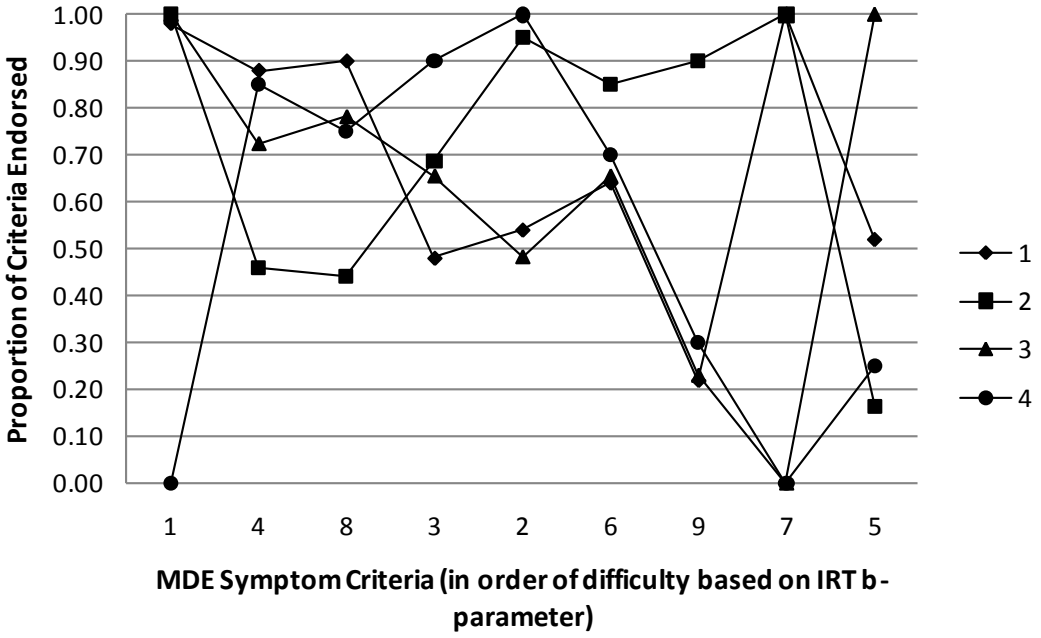

Figure 10. The means (proportions) of the nine DSM-IV major depressive episode symptom criteria for each of the first four out of eight clusters from the k-means cluster analysis.

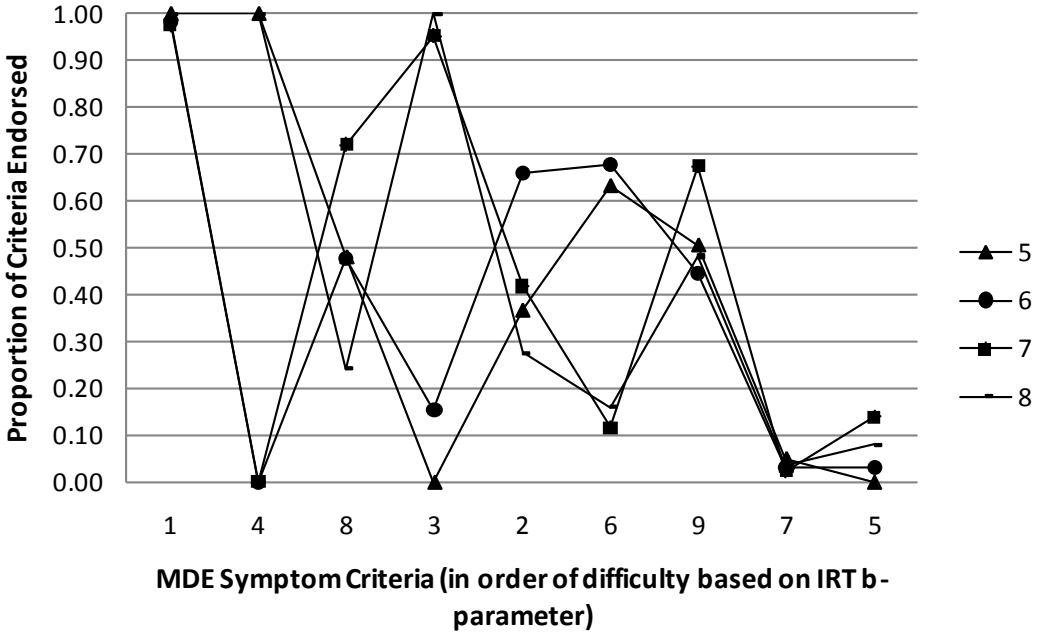

Figure 11. The means (proportions) of the nine DSM-IV major depressive episode symptom criteria for each of the second four out of eight clusters from the k-means cluster analysis.

CHAPTER V: DISCUSSION

Psychometric analysis: CTT

Central to all psychometric analysis is first ascertaining the dimensional structure of a set of indicator items (Anastasi & Urbina; Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton et al., 1991). A tetrachoric factor analysis on the nine MDE symptom criteria showed that there is one dominant factor among the criteria, which gives some evidence that the nine symptom criteria are indeed sampling some sort of unitary psychopathological construct. There appears to be no specific pattern in the loadings among the physical, emotional, vegetative, and cognitive symptoms of depression.

The item with the highest level of difficulty was depressed mood. The high level of difficulty for depressed mood is most likely an artifact of the selective NCS sampling procedure, which, as was discussed above, sampled individuals who already displayed at least some tendency toward depression.

A surprising finding for the symptom difficulty levels is that anhedonia, contrary to what might be expected given that it is one of the two gate symptoms, did not have either the first or second highest endorsement levels. Instead, sleep problems had the second highest level of difficulty. As noted above, anhedonia was part of a cluster of four symptoms that had endorsement rates in the low to mid .80's. Perhaps one reason that anhedonia did not stand out in its endorsement rate is that anhedonia is not directly associated with an explicit low mood, rather it is more akin to an absence of positive mood. As such, it may be more similar to two of the three other symptoms that had endorsement rates in the low to mid .80's (thinking/concentration problems and fatigue)

in that these symptoms all appear to be reflective of some sort of deficit in the capacity to carry out normal levels of functioning in various mental and/or emotional domains. In other words, anhedonia is decrease in the ability to feel a full normal range of emotions, thinking/concentration problems are a decrease in the ability to cogitate properly, and fatigue is a decrease in energy levels. Weight/appetite problems, which also had an endorsement rate in the low to mid .80's, do not quite fall in this category because they appear to be more of a dysregulation of food intake, and according to the wording of this particular symptom criterion, a person can either be eating too little or too much when depressed.

The remaining three symptoms, suicidal tendencies, psychomotor problems, and worthlessness/guilt, appear to be symptoms that are, like weight/appetite problems, dysregulation of normal process rather than a decrease in a particular level of ability from normal functioning. These results show that there is no consistent pattern among the physical, emotional, vegetative, and cognitive symptoms in terms of how they were endorsed by the individuals.

Psychometric analysis: IRT

To provide a more sophisticated psychometric analysis that goes beyond CTT, a series of IRT models was run: the 1PL, 2PL, 3PL, and reversed keyed 3PL models. The likelihood ratio tests for the 1PL, 2PL, and 3PL models showed that the 2PL model was the best fitting among the three standard IRT models and this was confirmed by the AIC, since it had lowest value among the three IRT models. The 2PL model was retained as the best fitting model.

Overall, the ICCs, item and test information curves, and the test characteristic curves for the 2PL model show that the DSM-IV symptom criteria for depression are good at providing the most information at a moderately low level on the latent trait spectrum for depression. The region in which the nine symptom criteria provide the most information corresponds to a symptom count of 4.5, which is close to the cutpoint of five symptoms for a diagnosis of depression.

The results of the 2PL model show that depressed mood has an extremely low difficulty parameter, which indicates that individuals who present themselves to a clinician as potentially having some sort of mood disorder will most likely be experiencing some sort of depressed mood before any other symptom criteria. Thus, it appears that having depressed mood as a gate criterion is appropriate. However, the other gate criterion, anhedonia, had a much higher difficulty level than depressed mood. Future research should examine the utility of having anhedonia as a gate criterion given its much higher difficulty level.

The 3PL and reversed keyed 3PL models were also examined for insights concerning the performance of the nine diagnostic criteria at the extreme ends of the latent spectrum for depression. Even though the 3PL model was not necessarily the best fitting IRT model (though it did appear to be the second best fitting IRT model after the 2PL model), its results are suggestive of some diagnostic biases that are present in the clinical screening process for depression.

The 3PL IRT model indicates that, except for all but the two most difficult symptoms, there may be a propensity for individuals with low to nonexistent levels of depression to over report depression symptoms, i.e., to report depression symptoms when

they are not expected to possess such symptoms. The results of the 3PL model suggest that for some reason there may be a bias for individuals to over report depressive symptoms in a clinical psychiatric interview setting. This bias may be due to, for instance, cognitive dissonance arising in individuals during a clinical interview who believe that since they are already undergoing a clinical psychiatric interview they should mention the presence of at least some psychiatric symptoms. Another possible cause for this bias may be that the mere mention of various clinical psychiatric symptoms elicits a social cognitive schema that decreases the threshold for an individual's personal judgment concerning the presence of a given psychiatric symptom. This is clearly an area for more future research.

The reversed keyed 3PL model was the best fitting among all the reversed keyed 3PL models. The reversed keyed 3PL model did not show any strong overall bias toward underreporting symptoms at the high end of the latent depressive spectrum, with the possible exception of psychomotor problems and fatigue.

Consistency/scalability analysis

The marginal posterior probability value for each individual's set of symptoms that was computed during the IRT scoring procedure was used as a measure of the consistency of each individual's pattern of symptom responding. The marginal posterior probability value for each individual was plotted against total score and a strong positive relationship between the marginal posterior probability value and the summed total score of the nine symptoms was found. In other words, as the number of symptoms increased for an individual, there was a greater tendency for an individual to have a pattern of

symptoms such that all the symptoms they possessed had difficulty levels up to the theta level of the individual.

Cluster Analysis

A cluster analysis was carried out on the pattern of symptom responding for individuals with the lowest 25% of marginal posterior probability values. An examination of the final solution (Figures 10 and 11) revealed some overall broad generalities about the pattern of symptom responding for individuals whose symptom patterns were judged to be relatively inconsistent.

The cluster analysis identified eight clusters with unique symptom patterns that were aberrant as compared to what would be expected based on the 2PL IRT item difficulty parameters and thus led to lower marginal posterior probability values for certain individuals. The eight clusters can be broadly divided into three kinds of groups, mostly based on three kinds of salient differences in the patterns of endorsement rates of the nine symptom criteria: first, a sharp difference in the rate of endorsement of the gate criteria of depressed mood and anhendonia, second, unexpectedly high rates of endorsement of one or more of the three most difficult symptoms (suicidal tendencies, worthlessness/guilt & psychomotor problems), and, third, varying levels of endorsement among the five symptoms with moderately low levels of difficulty (sleep problems, thinking/concentration problems, weight/appetite problems, anhedonia, and fatigue) even though they had almost identical ICCs.

One of the groups contained only one cluster (Cluster #4). This group was mainly characterized by 0% of the individuals in its sole cluster endorsing the gate criteria of depressed mood and 100% of the individuals in its cluster endorsing the gate criteria of

anhedonia. The five symptoms with moderately low difficulty levels had high rates of endorsement, as expected. The symptoms with high difficulty levels had low levels of endorsement, as expected. The pattern of endorsement of the symptoms of Cluster #4 indicates that individuals who do not have the gate criteria symptom of depressed mood but instead have the other gate criteria of anhedonia demonstrate a pattern of endorsement of the remainder of the seven symptoms that is as expected based on the difficulty levels of the symptoms.

The second group contained three clusters (Clusters # 1-3). All clusters in this group had 100% endorsement of the gate criteria of depressed mood and had unexpectedly high levels of endorsement of one or two of the three most difficult symptoms. Cluster #2 had unexpectedly high levels of endorsement of symptoms # 9 and 7. Cluster #3 had an unexpectedly high level of endorsement of symptom #5. Cluster #1 had unexpectedly high levels of endorsement of symptoms # 7 and 5.

A potentially clinically interesting set of associations concerning suicide and the other diagnostic criteria appear in Clusters #1-3. Suicide ideation is a difficult symptom and therefore it should appear with far less frequency in the population. However, in Cluster # 2, suicide ideation had an endorsement rate of approximately 90%. The pattern of endorsement rates for Cluster #2 indicate that individuals who have both depressed mood and anhendonia, as well as a lack of psychomotor retardation, are likely to be at high risk for suicide. These individuals also suffer from worthlessness/guilt and fatigue. What is particularly troubling from a clinical standpoint about individuals in Cluster #2 is that because they do not have psychomotor retardation, i.e., they are behaviorally active, they may have enough energy to actually carry out their suicidal tendencies. In both

Clusters # 1 and 3, individuals had lower rates of anhedonia and had concomitantly lower rates of suicide ideation.

The third group contained four clusters (Clusters # 5-8). All clusters in this group had 100% endorsement of the gate criteria of depressed mood and displayed, as expected, low levels of endorsement of the three most difficult symptoms. The differences among the four clusters in this group were due to the varying levels of endorsement of the five symptoms that have similar difficulty levels.

Interpretation of results from a clinical/categorical construct perspective

The results of the IRT analyses show that the DSM-IV symptom criteria of depression do not work well if depression is hypothesized as a continuum. Part of the problem with using the symptom criteria as indicators of a continuous construct is that they are not adequately distributed across the full range of the latent continuum of depression. However, in light of a clinical perspective that values a categorical conceptualization of depression, the implications of the results of the IRT analysis take on a different meaning. For purposes of diagnosis, the symptom criteria for depression do appear to be efficient in indentifying individuals with a potential need of treatment. In order to achieve a diagnosis of depression, an individual needs either to have one of the two gate symptoms, depressed mood or anhedonia. Depressed mood is a relatively easy diagnostic criteria and many individuals in the sample possess this trait. Once a gate criterion has been identified, which most likely would be depressed mood, individuals then need to have four additional symptoms. Of the remaining eight symptoms beside depressed mood, five have a somewhat more moderate level of difficulty on the latent continuum, which insures that there will be a greater likelihood of identifying depressed

individuals than if those symptoms had higher difficulty levels. Furthermore, the 3PL model shows that six of the nine diagnostic criteria have large *c* parameters. The inflated *c* parameters may be an indication that individuals who present a depressed mood to a diagnosing clinician have a bias to over report or over inflate their experience of the six diagnostic criteria with the high *c* parameters. Again, this may not be problematic if the goal is to identify as many individuals as possible who are truly depressed.

Evaluation of cutpoint of minimum of 5 symptoms

The results of the IRT analysis help bring a "conceptual order" to the findings of the literature reviewed above on the problems associated with the cutpoint of five symptoms for a diagnosis of depression and the psychiatric impairments associated with the subthreshold depressive disorders that are defined by using a less restrictive cutpoint. First, the results of the IRT analyses suggest that this cutpoint criterion may be statistically somewhat arbitrary. The 2PL IRT model shows that five of the symptoms of depression (sleep problems, thinking/concentration problems, anhedonia, weight/appetite problems, fatigue) have similar ICCs with levels of difficulty in the approximate theta range of -2.3 to -1.7. This pattern of difficulty parameters from the 2PL model indicates that even the presence of one of these symptoms is stochastically indicative of an already moderate level of depression that is more severe than the presence of depressed mood alone. Furthermore, an individual with only two or three symptoms from this cluster of five psychometrically similar symptoms will likely have the same level of depression as an individual with four or five of these five symptoms. In such a case, however, the individual with two or three of these symptoms will not be diagnosed as depressed even if his or her level of depression falls in approximately the same range on the latent

continuum as an individual with four or five of these symptoms. As the number of symptoms increases beyond a cutpoint of five, then individuals statistically are most likely to start having one or more of the three most difficult symptoms (suicidal tendencies, psychomotor problems, and worthlessness/guilt), which of course indicates the presence of even greater psychiatric distress. Thus, the cutpoint of five symptoms may be somewhat too severe for purposes of detecting individuals with depression.

The results of the IRT analysis also give a psychometric validation to Kramer's (2005) conclusion that "symptom-free recovery" (p. 164) is necessary for a patient with depression in treatment. If an individual has five or more symptoms and then decreases down to, say, only two or three symptoms, it is obvious that he or she is most likely located at a point on the latent scale of depression that corresponds to at least a moderate level of depression, which is still associated with some sort of disability in normal psychological functioning as evidenced by the *b* parameters for the criteria of sleep problems, thinking/concentration problems, anhedonia, weight/appetite problems, and fatigue. From the perspective of the IRT model, a patient should at most have only the criterion of depressed mood in order to feel well enough that they are not incapacitated in any way in their day to day functioning. In the case of the presence of only depressed mood, the individual may simply be experiencing sadness.

In the above reviewed literature on subthreshold depression, the different subcategories of depression were shown to have clinical significance. One obvious solution to the problem of proliferating subthrehold categories of depression is to simply do a symptom count. On the surface, a total score symptom count may appear to be a better alternative to the cutpoint rule. However, the IRT analyses in the current study

show that this would be problematic for the current set of DSM-IV diagnostic criteria for depression. The ICCs for all items are not distributed evenly across the entire range of the latent spectrum. The 2PL IRT model shows that some of the symptoms have lower discrimination parameters than others, which indicates that symptoms should not be weighted equally for a diagnosis of depression. For a total score approach toward diagnosing and reporting depression to be workable in a clinical setting, the symptoms would need to be revised and/or edited such that they have better discrimination parameters and a wider range of difficulty parameters.

Mapping symptoms onto the latent continuum of depression

A useful feature of IRT is that it allows observed indicators to be mapped onto different levels of a construct and thus, the indicators as well as persons can be placed on a common scale (Embretson & Reise, 2000). In the case of the MDE symptom criteria, this allows for the individual symptoms to be mapped onto specific levels of depression, which conversely can also be used to predict which symptoms are most likely to arise at different levels of depression. Using the ICC profiles, it is tempting to try to infer a developmental sequence of depression symptoms, i.e., which symptoms are most likely to be experienced at the beginning of the disorder, which presumably is associated with lower levels of depressive mood, and then which other symptoms are most likely to appear with increasing severity of depressive mood. As noted above, the most prevalent symptom is depressed mood. After depressed mood, there is a cluster of five symptoms (sleep problems, thinking/concentration problems, anhedonia, weight/appetite problems, fatigue) that appear most likely to manifest. It appears that suicidal tendencies develop later, though there is a high degree of variability of when suicidal tendencies may appear

during the development of depression as indicated by the lower *a* parameters in the 2 and 3PL models and the high *c* parameter in the 3PL model for the symptom of suicidal tendencies. After suicidal tendencies, it does appear as if individuals with increasing levels of depression experience a "shutdown" of their behaviors and cognitions as reflected by the symptoms of psychomotor problems and worthlessness/guilt. Of course, this inferred developmental model of depression assumes that individuals start off with a low level of depression and then progress to higher levels of depression, as defined by the latent IRT continuum. It is possible that individuals can simply start off a depressive episode at an already moderate to high level of depression, in which case many symptoms would then appear all at once.

Implications of IRT results for future diagnostic rules for depression

From the clinical/categorical perspective, the results of the IRT analysis suggest that the format of the current DSM-IV diagnostic criteria for a MDE could remain as is for future versions of DSM-IV and still be viable for its stated purpose of identifying individuals who are depressed. However, the IRT analyses do also suggest that there is at least some room for the improvement and fine tuning of the diagnostic rules for depression.

One obvious possibility is that perhaps the diagnostic criteria could be rewritten and/or expanded with new symptoms so that they have a better range of coverage across the full latent scale of depression. Criteria with a compound format could be disaggregated, e.g., asking separately whether there is a psychomotor agitation or psychomotor retardation. However, given the preference for the use of the cutpoint technique by the creators of DSM in order to create a categorical diagnostic system out of

a list of symptom criteria and the potential reluctance of clinicians to switch to a dimensional system of diagnosis, this would probably in the end not be a useful approach since no matter what kind of diagnostic symptom criteria are used to define the diagnosis of depression, any increase in the specificity of symptom wording would eventually be eliminated by the use of the cutpoint technique. Also, properly pilot testing a new set of symptom criteria in a psychometric study in order to determine their psychometric parameters would probably be an expensive proposition and a project into which the creators of the next edition of DSM may not want to invest resources. From a pragmatic point of view, it would be best to improve the diagnostic criteria for a MDE keeping a format similar to the one currently found in DSM-IV. While this may not be optimal from a strict psychometric theory perspective, it does have the advantage of retaining the categorical definition of depression with which clinicians are familiar and it would involve minimal cost.

The current IRT analyses do provide some rich fodder for speculation on how to improve the current DSM-IV diagnostic rules for depression. In particular, the diagnostic criterion of suicidal tendencies stands out as a symptom that deserves special attention.

Suicidal tendencies is a more difficult symptom, which suggests that individuals may not initially become suicidal when symptoms of depression first present themselves. Also, in the 2PL model, suicidal tendencies has the smallest *a* parameter, which suggests that it does not do as good a job discriminating individuals with higher levels of depression as the other symptom criteria. In other words, the relationship between depression and suicide is not as strong as the relationships between the latent trait of depression and the other symptoms. The lower *a* parameter from the 2PL model as well

as the high *c* parameter from the 3PL model for suicidal tendencies also suggest that suicidal tendencies can become a problem at lower levels of depression despite the symptom's relatively high *b* parameter.

The lower IRT *a* parameter for suicidal tendencies may in part be due to suicide as an implicit risk factor for a host of other nondepressive psychiatric conditions that are comorbid with depression such as anxiety disorders, schizophrenia and personality disorders (APA, 2000; Barlow & Durand, 2005). It is quite likely that if symptom criteria from various different psychiatric disorders were simultaneously assessed, suicide would most likely show a high degree of statistical and psychometric multidimensionality. Thus, in a future DSM diagnostic nosological scheme, perhaps suicide would most profitably be considered as a special symptom criterion that would be specially probed for by clinicians, independent of the presence or absence of the other symptom criteria for depression.

The recommendation to put suicidal tendencies in its own special category leaves a remaining set of eight symptoms from the list of current MDE symptoms. The 2PL IRT model shows that five of these remaining eight symptoms have highly similar ICC's that cluster together on the latent scale of depression at a location that corresponds to a moderate level of depression. Thus, the probability is quite high that individuals with a level of depression around the theta level of approximately -.5 will show one or more of these five symptoms. In a future edition of DSM, these five symptoms could be used to form a "core" set of depression criteria in the diagnostic rules for depression, given the similarity of their stochastic profile.

The above proposals would then leave depressed mood as the only gate criteria required for a diagnosis of depression. Anhedonia, which currently in DSM-IV is a gate symptom, is a more difficult symptom than depressed mood and it has an ICC that makes it clearly a member of the revised "core" set of five symptom criteria described above. However, the implications of removing anhedonia as a gate criterion need to be empirically assessed. Also, the symptoms of psychomotor problems and worthlessness/guilt would not be part of the revised "core" set of symptoms. Future research needs to assess whether it would be better to include them in the list of "core" symptoms or place them in a separate specifier category.

Finally, the cutpoint for the "core" symptom set could be lowered to three symptoms, which with the addition of depressed mood as a gate criterion, would lead to a requirement of a total of four symptoms for a diagnosis of depression. The current cuttpoint of five or more symptoms, with at least one of them being depressed mood and anhedonia, may be somewhat too conservative. However, this needs to be empirically investigated in future research.

Implications of IRT results for depression from a continuous perspective

Instead of conceptualizing depression as a categorical disease entity, an alternative way to conceptualize depression is as a continuum. The results of the current study show that, from a purely psychometric perspective, the current set of DSM-IV diagnostic criteria for a MDE is inefficient as a measure of depression if the construct of depression is conceptualized as a continuum. This is primarily due to the incomplete coverage of the symptom criteria across the full range of the depressive spectrum. The results of the IRT analysis therefore also imply that the DSM-IV symptom criteria for a

MDE should not be used in any situation that requires the use of a continuous scale of depression, such as an outcome measure in a clinical trial of a treatment for depression where the level of depression needs to be tracked longitudinally. In such research settings, there is a need to capture levels of depression on a more fine grain level of resolution than a simple dichotomy. A good psychometric instrument for testing depression in such a situation should have a mix of easy, moderate and hard diagnostic indicators.

However, because the clinical psychiatric community may be reluctant to adopt a continuous model of depression for diagnostic purposes (APA, 2000), for the foreseeable future, a de facto compromise in measuring depression may have to be adopted. The DSM-IV diagnostic criteria for a MDE, and it's no doubt categorical based successors in future editions of DSM, will probably continue to be the "official" definition of depression for diagnostic purposes, while psychometric instruments such as the Beck Depression Inventory (Beck et al., 1961) or Hamilton Depression Scale (Hamilton, 1967) will probably be the best choice for use as continuous outcome measures for depression research.

Future Research

IRT offers the possibility for more sophisticated future research on the diagnostic criteria for depression than is possible with CTT alone. One area of research for which IRT would be useful is examining an expanded and/or disaggregated list of core symptom criteria. Currently, there are some symptom criteria that are relegated to the specifer section for depression. For example, "loss of pleasure in all, or almost all, activities" (APA, 2000, p. 420) and "lack of reactivity to usually pleasurable stimuli"

(APA, 2000, p. 420) are unique symptoms for the specifer of melancholic depression. "Leaden paralysis" (APA, 2000, p. 422) and "long-standing pattern of interpersonal rejection sensitivity" (APA, 2000, p. 422) are unique symptoms for the specifier of atypical features depression. IRT would be useful to determine how these unique symptom criteria for specifiers of depression relate to the overall latent continuum of depression and also help determine if these symptoms should be incorporated into the core list of symptom criteria for depression. IRT could also be used to compare certain compound core symptom criteria with their disaggregated versions. This kind of analysis would be useful to determine whether the disaggregated versions of compound symptoms provide similar or different information on the level of depression for an individual as compared to the original compound criteria. IRT lends itself well to an analysis of the differential item functioning (DIF) of the symptom criteria, which in this case would be useful for determining how the different symptom criteria perform in different population subgroups, e.g., sex, socioeconomic status, and contexts, e.g., whether the nongate symptoms perform differently among individuals who meet the gate criteria for depression.

DSM-IV is essentially considered a "gold standard" for definitions of various kinds of psychopathology in North America. However, as noted above, it is not useful as an outcome measure of depression for research purposes. Instead, inventories such as the Beck Depression Inventory (Beck et al., 1961) and the Hamilton Depression Scale (Hamilton, 1967) are often used instead in research. An important study that could only be conducted using IRT would be a linkage/equating study in which items on the Beck Depression Inventory and Hamilton Depression Scale would be placed on the same

common scale along with the DSM-IV symptom criteria of depression. Such a study would determine how these inventories function with respect to the "official" criteria of depression. One advantage of such a study is that once the inventories are mapped onto a common scale with the DSM-IV criteria, better cutpoints for the inventories can be created. Also, such a linking between the DSM criteria and commonly used inventories is that it would lead to a better understanding of how the items from a particular inventory relate to the actual "official" DSM diagnosis of depression.

Conclusion

The current study conducted a CTT and IRT psychometric analysis of the nine DSM-IV symptom criteria for a MDE. Overall, it does appear that the MDE symptom criteria are useful as a screening tool for the presence of depression. The pattern of the symptom ICCs indicates that, for the most part, the symptom criteria are useful for capturing individuals who have a moderate level of depression. However, the IRT analyses have revealed several important aspects concerning these symptom criteria. First, they are not useful as a continuous measure of depression because the difficulty parameters of the symptom criteria are not spread out enough across the full latent spectrum of depression. Second, the symptom criteria can be improved upon by eliminating three of the symptom criteria and placing them into specifier categories.

The DSM-IV definition of a MDE is based on a categorical approach toward mental illness. The categorical approach in this case is maintained through the imposition of a cutpoint of five symptoms on the list of symptoms. The IRT analyses have shown that a cutpoint of five symptoms required for a diagnosis of depression is most likely too conservative. Perhaps a cutpoint of two or three symptoms would be adequate. While a

continuous approach toward depression may be more useful in a variety of different research settings, the IRT analyses have shown that the categorical approach toward depression found in DSM-IV is useful in clinical settings for identifying individuals in need of treatment.

References

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychol Med, 35*(4), 475-487.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

American Psychiatric Association (APA). (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

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

American Psychiatric Association (APA). (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association (APA). (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Angst, J., & Merikangas, K. (1997). The depressive spectrum: diagnostic classification and course. *Journal of Affective Disorders, 45,* 31-40.

Baker, F. B. (2001). *The basics of item response theory*. Retrieved January 2007, from http://edres.org/irt/.

Barlow, D. H., & Durand, V. M. (2005). *Abnormal psychology: An integrative approach* (4th ed.). Blemont, CA: Thomson Wadsworth.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314.

Buchwald, A. M., & Rudick-Davis, D. (1993). The symptoms of major depression. *Journal of Abnormal Psychology, 102(2),* 197-205.

Carroll, B. J. (1984). Problems with diagnostic criteria for depression. *Journal of clinical psychiatry, 45(7, Section 2),* 14-18.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88-106.

Christensen, L. B. (1988). *Experimental methodology* (4th ed.). Newton, MA: Allyn and Bacon.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Reviews of Psychology, 46,* 121-153.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment, 68*(3), 508-531.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11 Suppl 3), S50-59.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT* [Computer software manual]. Lincolnwood, IL: Scientific Software International, Inc.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143-165.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24*(2), 133-148.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*(3), 201-212.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). New York: Halsted Press.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Faravelli, C., Servi, P., Arends, J. A., & Strik, W. K. (1996). Number of symptoms, quantifications, and qualification of depression. *Comprehensive Psychiatry, 37(5),* 307-315.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, (26),* 57-63.

Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality, 31*(4), 439-485.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *J Pers Soc Psychol, 78*(2), 350-365.

Frances, A., Pincus, H. A., Widiger, T. A., Davis, W. W., & First, M. B. (1990). DSM-IV: Work in progress. *American Journal of Psychiatry, 147(11),* 1439-1448.

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10*(4), 544-565.

Gordis, L. (2000). *Epidemiology* (2nd ed.). Philadelphia: W. B. Saunders.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532-560.

Gray-Little, B., Williams, V. S., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*(5), 443-451.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage Publications, Inc.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Clinical Psychology, 6(4),* 278-296.

Harkness, A. R., McNulty, J. L., & Ben-Porath, Y. S. (1995). The Personality Psychopathology Five (PSY-5): Constructs and MMPI-2 scales. *Psychological Assessment, 7*(1), 104-114.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(Suppl9), II28-II42.

Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology, 61*(6), 932-945.

Judd, L. L., Akiskal, H. S., & Paulus, M. P. (1997). The role and clinical significance of subsyndromal depressive symptoms (SSD) in unipolar major depressive disorder. *Journal of Affective Disorders, 45,* 5-18.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. van Duijn & T. A. Snijders (Eds.), *Essays on item response theory* (Vol. 157, pp. 247-276). New York: Springer-Verlag.

Kalton, G. (1983). *Introduction to survey sampling.* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035). Newbury Park, CA: Sage.

Kendler, K. S. (1990). Toward a scientific psychiatric nosology: Strengths and limiations. *47*, 969-973.

Kendler, K. S., & Gardner, C. O. (1998) Boundaries of major depression: An evaluation of DSM-IV criteria. *American Journal of Psychiatry, 155(2),* 172-177.

Kessler, R. C. (n.d. a) *National comorbidity survey: Replication (NCS-R), 2001-2003, Section 2: Screener* [questionnaire]. Retrieved January 2007, from http://www.hcp.med.harvard.edu/ncs/ftpdir/replication/US%20Screener.pdf

Kessler, R. C. (n.d. b) *National comorbidity survey: Replication (NCS-R), 2001-2003, Section 3: Depression* [questionnaire]. Retrieved January 2007, from http://www.hcp.med.harvard.edu/ncs/ftpdir/replication/US%20Depression.pdf

Kessler, R. C. (n.d. c) *NCS-R Screener notes to all users.* Retrieved January 2007, from http://www.hcp.med.harvard.edu/ncs/notes_depression.php

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.-E., Walters, E. E., Zaslavasky, A., & Zheng, H. (2004). The US National Comborbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research, 13(2),* 69-92.

Kessler, R. C., & Merikangas, K. R. (2004). The National Comorbidity Survey Replication (NCS-R): Background and aims. *International Journal of Methods in Psychiatric Research, 13(2),* 60-68.

Kessler, R. C., & Üstün, T. B. (2004). The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research, 13(2),* 93-121.

Kessler, R. C., Zhao, S., Blazer, D. G., & Swartz, M. (1997). Prevalence, correlates, and course of minor depression and major depression in the national comorbidity survey. *Journal of Affective Disorders, 45,* 19-30.

Kline, P. (1998). *The new psychometrics: Science, psychology and measurement*. Florence, KY: Taylor & Frances/Routledge.

Kramer, P. (2005). *Against depression*. New York: Penguin Books.

Krueger, R. F., & Piasecki, T. M. (2002). Toward a dimensional and psychometrically-informed approach to conceptualizing psychopathology. *Behaviour Research and Therapy, 40,* 485-499.

Langenbucher, J. W., Labouvie, E., Martin, C. S., Sanjuan, P. M., Bavly, L., Kirisci, L, & Chung, T. (2004). An application of item response theory analysis to alcohol,

cannabis, and cocaine criteria in DSM-IV. *Journal of Abnormal Psychology, 113(1),* 72-80.

Loevinger, J. (1993). Measurement of personality: True or false. *Psychological Inquiry, 4*(1), 1-16.

Maier, W., Gänsicke, M, & Weiffenbach, O. The relationship between major and subthreshold variants of unipolar depression. *Journal of Affective Disorders, 45,* 41-51.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*(2), 261-279.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354-368.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Brown (Eds.), *Test validity* (33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2),* 5-11.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague: Mouton & Co.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer-Verlag.

Muthen, L. K., & Muthen, B. O. (1998-2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthen & Muthen.

Nathan, P. E., & Langenbucher, J. W. (1999). Psychopathology: Description and classification. *Annual Reviews of Psychology, 50,* 79-107.

National Comorbidity Survey. (n.d.). *NCS: Answers to frequently asked questions.* Retrieved January 2007, from http://www.hcp.med.harvard.edu/ncs/faqncs.php

Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Application to smoking cessation. *Psychological Methods, 4*(2), 173-191.

Panter, A. T., Swygert, K. A., Grant Dahlstrom, W., & Tanaka, J. S. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68*(3), 561-589.

Parker, G. (2000). Classifying depression: Should paradigms lost be regained? *American Journal of Psychiatry, 157(8),* 1195-1203.

Parker, G. (2005). Beyond major depression. *Psychol Med, 35*, 467-474.

Philipp, M., Maier, W., & Delmo, C. D. (1991a). The concept of major depression: I. Descriptive comparison of six competing operational definitions including ICD-10 and DSM-III-R. *European Archives of Psychiatry and Clinical Neuroscience, 240,* 258-265.

Philipp, M., Maier, W., & Delmo, C. D. (1991b). The concept of major depression: II. Agreement between six competing operational definitions in 600 psychiatric

inpatients. *European Archives of Psychiatry and Clinical Neuroscience, 240,* 266-271.

Philipp, M., Maier, W., & Delmo, C. D. (1991c). The concept of major depression: III. Concurrent validity of six competing operational definitions for the clinical ICD-9 diagnosis. *European Archives of Psychiatry and Clinical Neuroscience, 240,* 272-278.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381-394). New York: Springer-Verlag.

Radloff, L. S. (1977). The CES-D Scale: a self-report depression scale for research in general population. *Applied Psychological Measurement, 1,* 385-401.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4(3),* 207-230.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. P. Reise (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219-241). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228-238.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184.

Reiser, M. (1989). An application of the item-response model to psychiatric epidemiology. *Sociological Methods and Research, 18*(1), 66-103.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer-Verlag.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*(2), 282-307.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66(1),* 63-84.

Sadek, N., & Bona, J. (2000). Subsyndromal symptomatic depression: A new concept. *Depression and anxiety, 12,* 30-39.

SAS Institute Inc. (2004). *SAS OnlineDoc® 9.1.3* [Electronic computer software manual]. Cary, NC: Author.

Shors, T. J., & Leuner, B. (2003). Estrogen-mediated effects on depression and memory formation in females. *Journal of Affective Disorders, 74*(1), 85-96.

Sijtsma, K., & Molennar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sptizer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry, 35,* 773-782.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). New York: Allyn & Bacon.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Medical Care, 44*(11 Suppl 3), S152-170.

van Praag, H. M. (2001). The diagnosis of depression in disorder. *Australian and New Zealand Journal of Psychiatry, 32*, 767-772.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64*(3), 545-576.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 49-79). Palo Alto, CA: Davis-Black.

Widiger, T. A., & Trull, T. J. (1991). Diagnosis and clinical assessment. *Annual Reviews of Psychology, 42*, 109-133.

Wilson, M., Allen, D. D., & Li, J. C. (2006a). Improving measurement in health education and health behavior research using item response modeling: Comparison with the classical test theory approach. *Health Educ Res, 21 Suppl 1*, i19-i32.

Wilson, M., Allen, D. D., & Li, J. C. (2006b). Improving measurement in health education and health behavior research using item response modeling: Introducing item response modeling. *Health Educ Res, 21 Suppl 1*, i4-18.

Wittchen, H.-U., Knauper, B., & Kessler, R. C. (1994). Lifetime risk of depression. *British Journal of Psychiatry, 165*(Suppl 26), 16-22.

Zimmerman, M., McGlinchey, J. B., Young, D., & Chelminski, I. (2006a). Diagnosing major depressive disorder I: A psychometric evaluation of the DSM-IV symptom criteria. *Journal of Nervous and Mental Disease, 194*(3), 158-163.

Zimmerman, M., McGlinchey, J. B., Young, D., & Chelminski, I. (2006b). Diagnosing major depressive disorder introduction: An examination of the DSM-IV diagnostic criteria. *Journal of Nervous and Mental Disease, 194*(3), 151-154.

Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57*(6), 963-969.

Appendix

Table 10
1-PL IRT Parameters for the Major Depressive Episode Symptom Criteria

Symptom Criteria	parameters			
	a	*a* SE	*b*	*b* SE
1. Depressed Mood	0.675	0.024	-4.427	0.189
2. Anhedonia	0.675	0.024	-1.782	0.066
3. Weight/Appetite Problems	0.675	0.024	-1.809	0.065
4. Sleep Problems	0.675	0.024	-2.323	0.076
5. Psychomotor Problems	0.675	0.024	0.046	0.051
6. Fatigue	0.675	0.024	-1.67	0.063
7. Worthlessness/Guilt	0.675	0.024	0.373	0.051
8. Thinking / Concetration Problems	0.675	0.024	-2.026	0.072
9. Suicidal Tendencies	0.675	0.024	-0.813	0.053

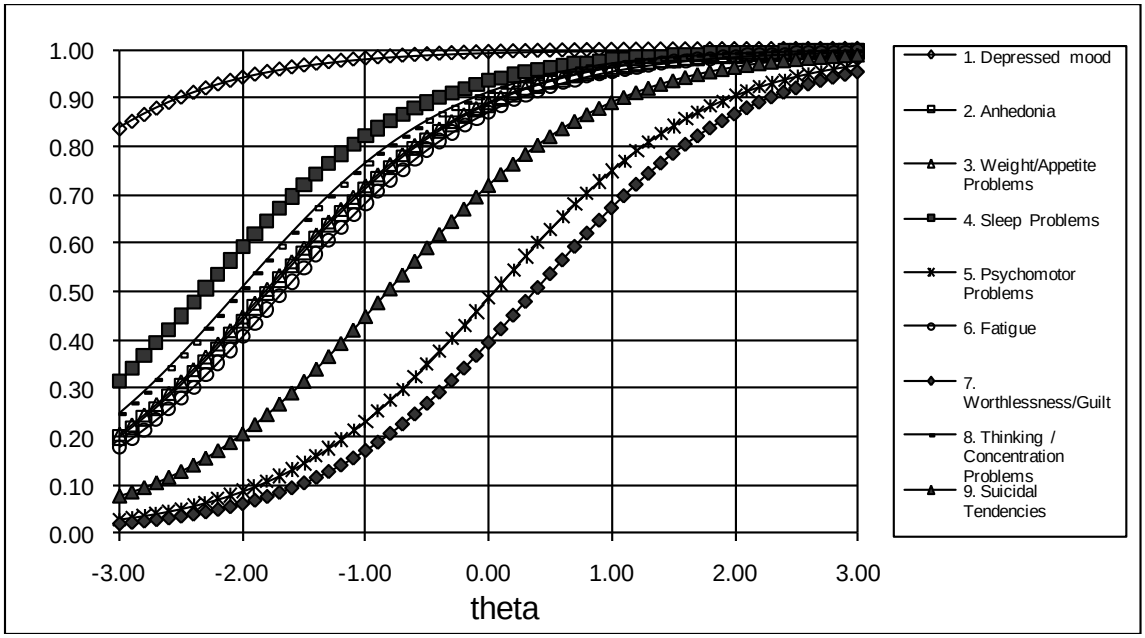

Figure 12. Item characteristic curves for the 1PL IRT model.

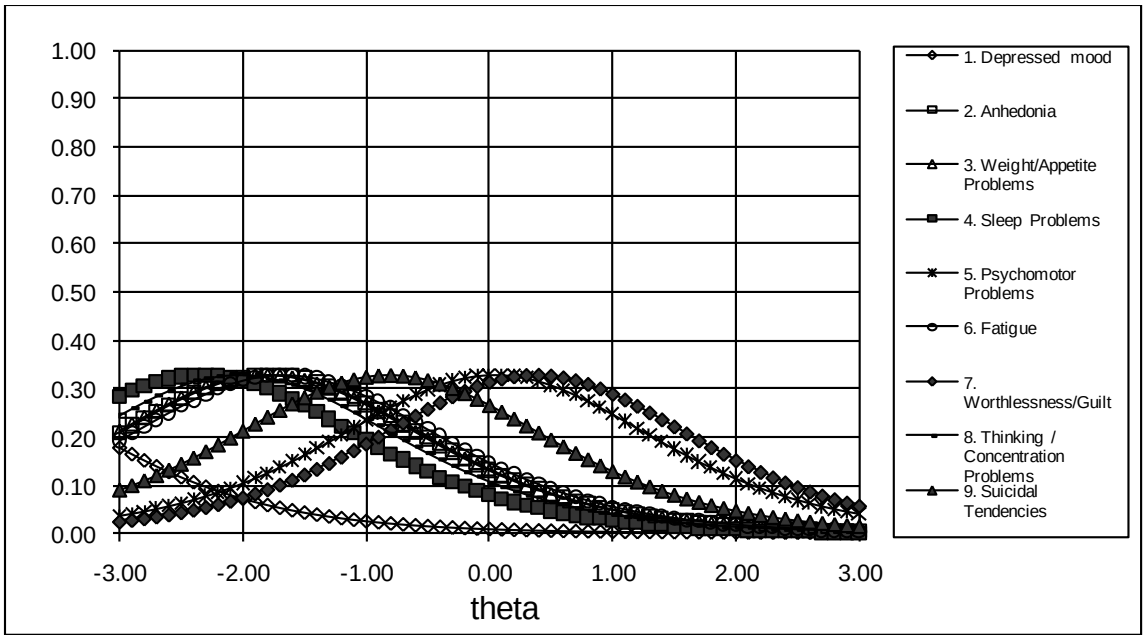

Figure 13. Item information curves for 1PL IRT model.

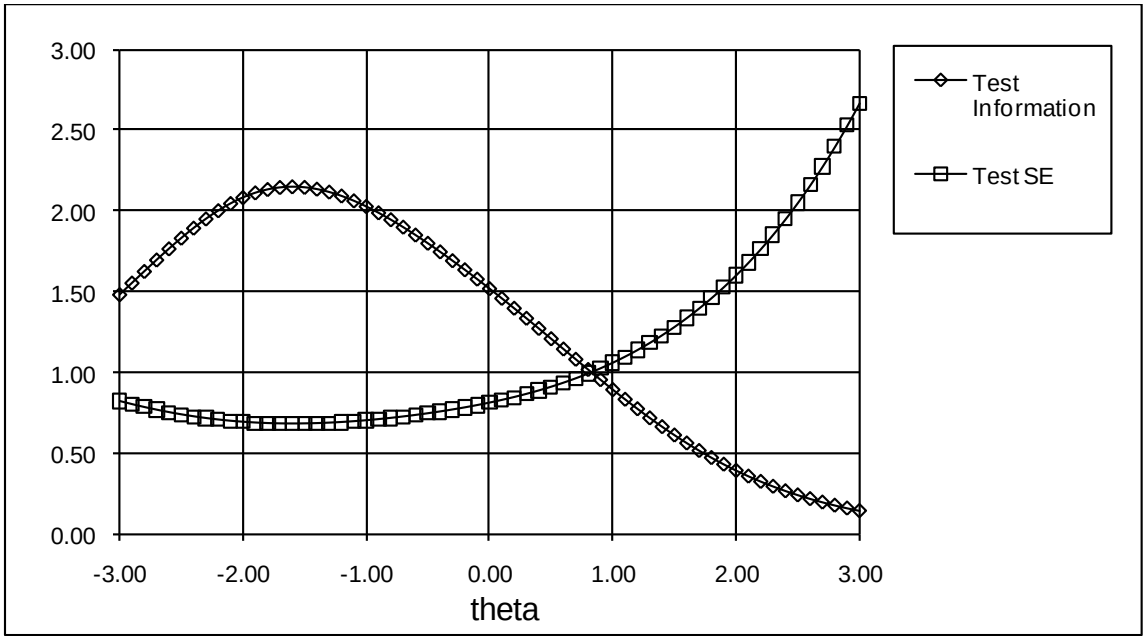

Figure 14. Test information and standard error curves for 1PL IRT model.

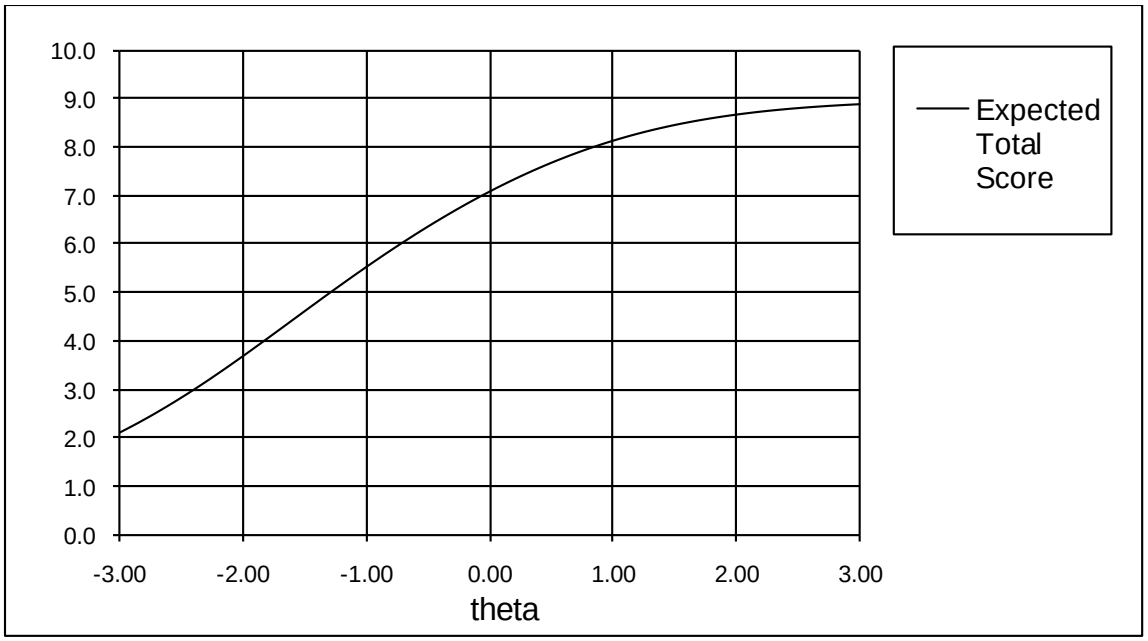

Figure 15. Test characteristic curve for 1PL IRT model.

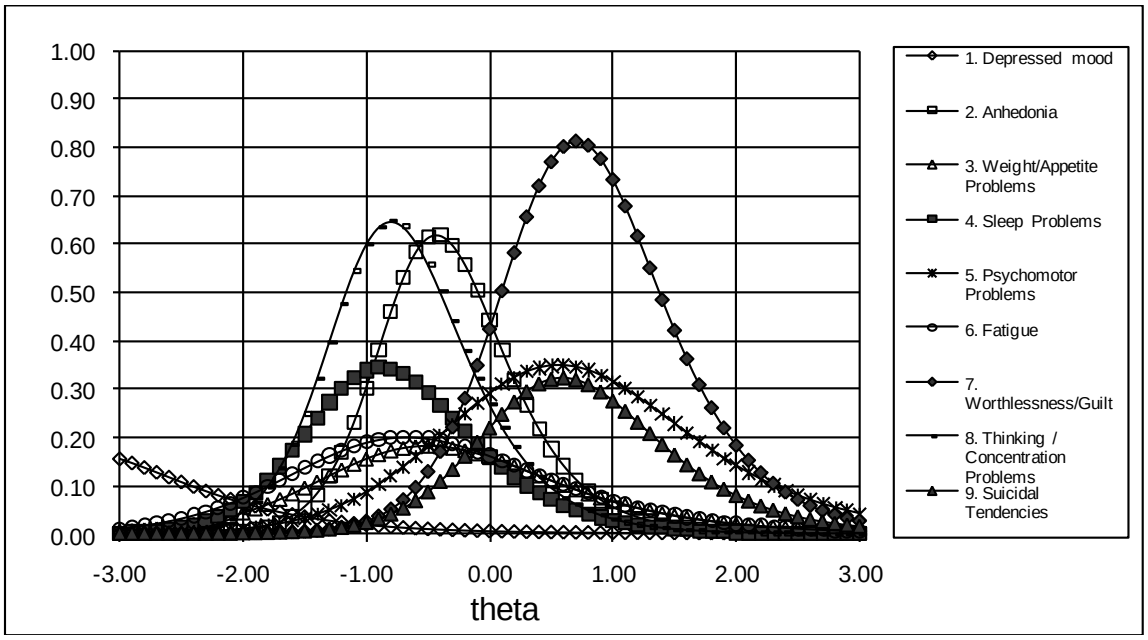

Figure 16. Item information curves for 3PL IRT model.

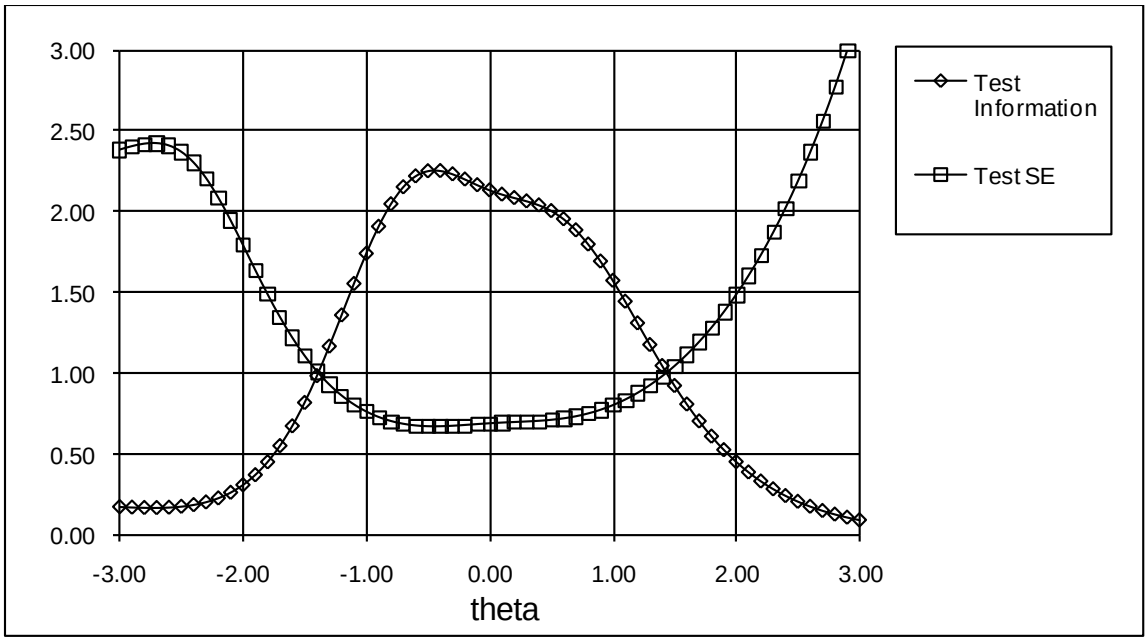

Figure 17. Test information and standard error curves for 3PL IRT model.

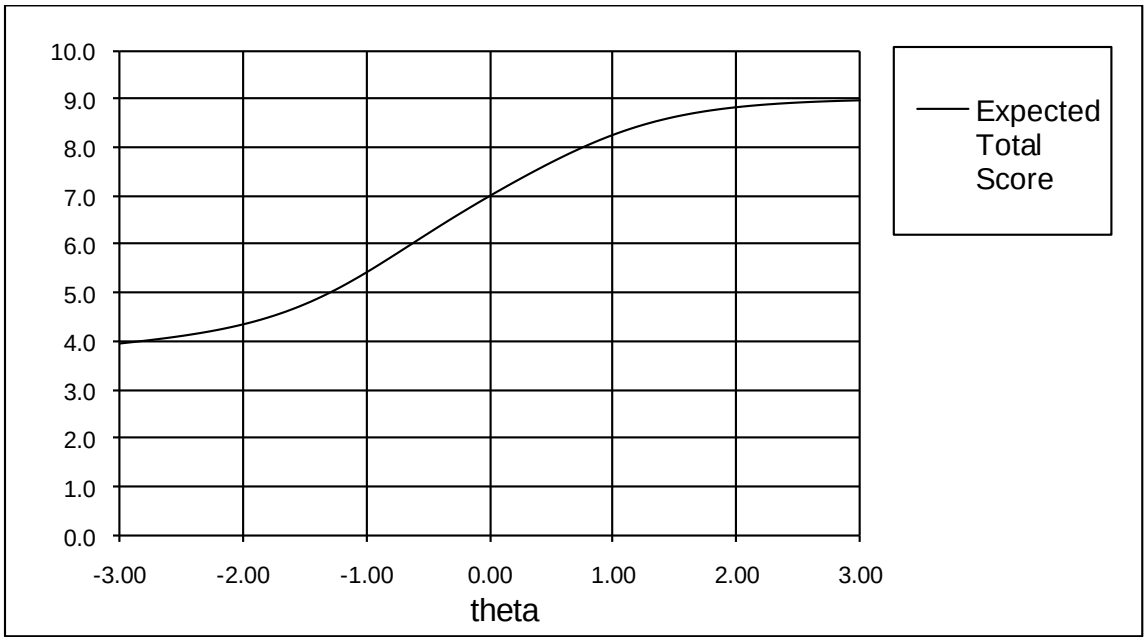

Figure 18. Test characteristic curve for 3PL IRT model.

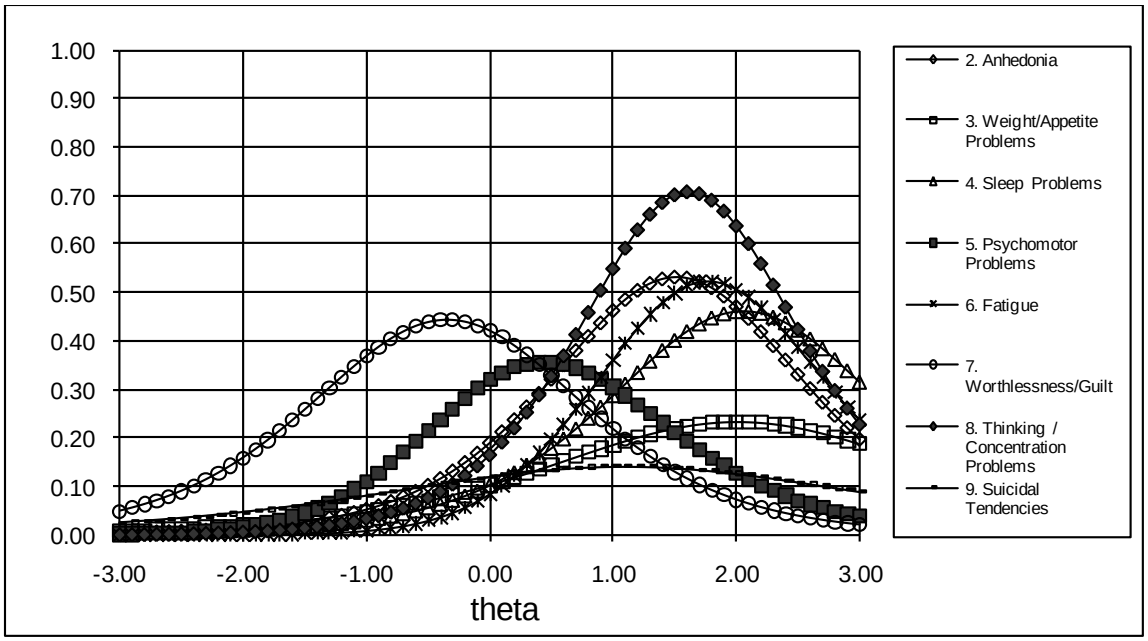

Figure 19. Item information curves for reversed key 3PL IRT model.

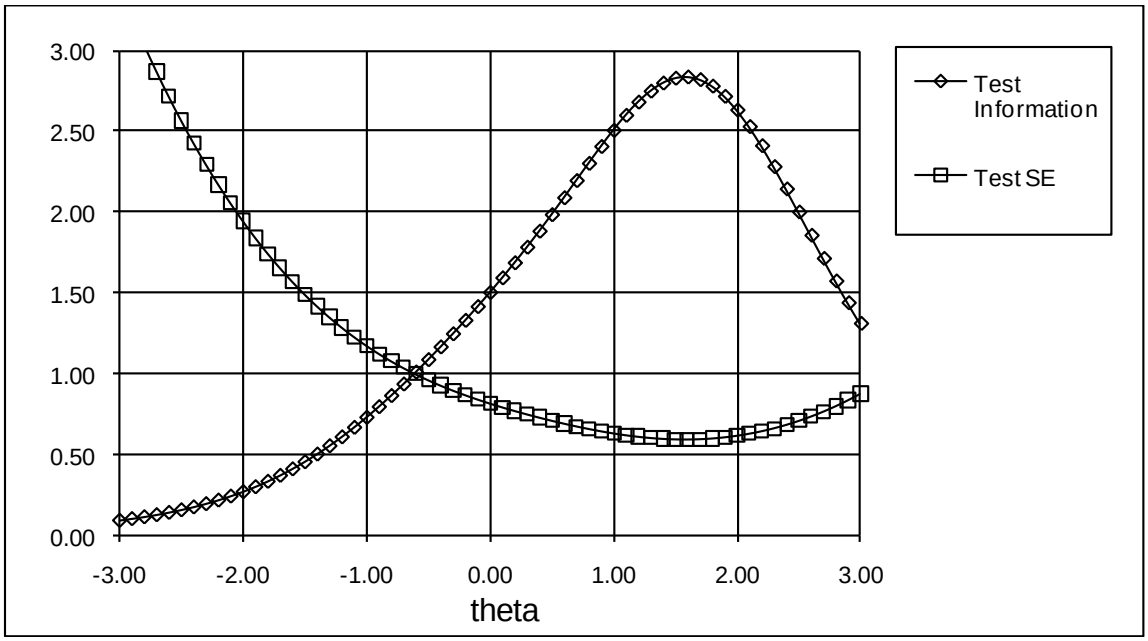

Figure 20. Test information and standard error curves for reversed key 3PL IRT model.

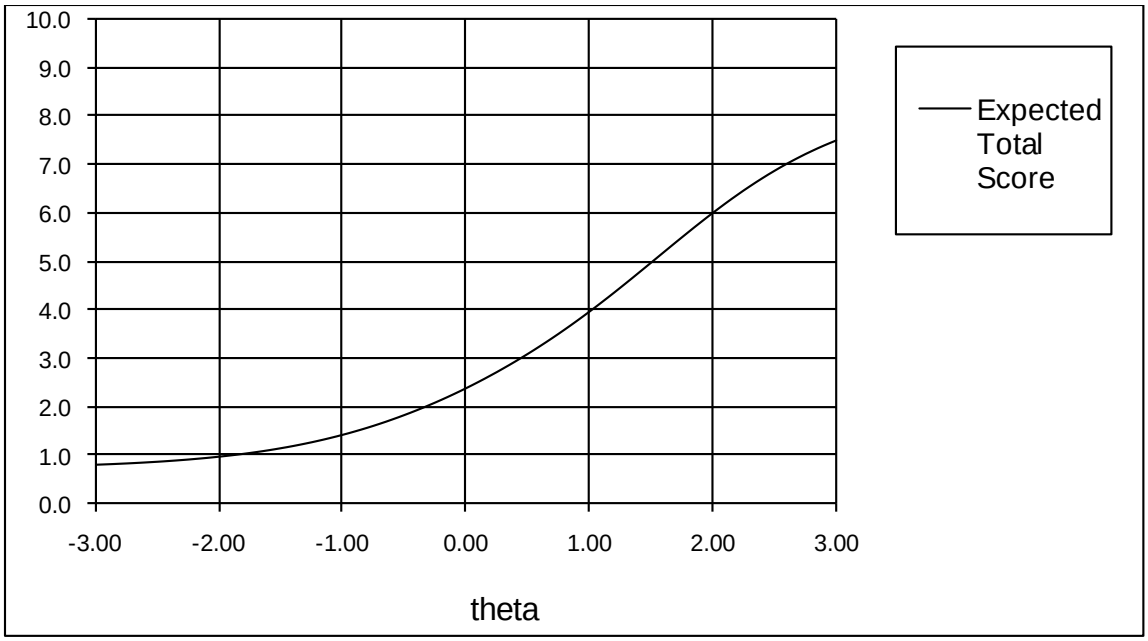

Figure 21. Test characteristic curve for reversed key 3PL IRT model.

CURRICULUM VITAE

Anthony P. Pawlak

EDUCATION

May 2010 — Rutgers University
New Brunswick, NJ
Ph.D. in Education; Concentration in Applied Statistics and Psychometrics
Dissertation Title: *A Classical Test Theory and Item Response Theory Analysis of the DSM-IV Symptom Criteria for a Major Depressive Episode Using Data from the National Comorbidity Survey – Replication.*

October 2004 — Rutgers University
New Brunswick, NJ
M.S. in Psychology; Biopsychology and Behavioral Neuroscience Program
Master's Title: *Dose Dependent Effects of Cocaine on the Firing of Rat Striatal Neurons Related to Head Movement in an Operant Head Movement Task Depend on Predrug Firing Rate.*

May 1996 — Franklin & Marshall College
Lancaster, PA
B.A., Cum Laude, in Psychology

WORK EXPERIENCE

Fall 2005 to present — Consultant
New Brunswick, NJ
Statistics Consulting
- Advised on data analysis, application of psychometric theory, and research method design for four separate research projects.
- Implemented and prepared the results of structural equation modeling, classical test theory analysis, hierarchical linear modeling and time-series analysis for clients.

Fall 2001 to Spring 2002 — Educational Testing Service
Princeton, NJ
Graduate Intern, *Analysis of SAT Data*
- Modeled and graphed SAT item data using SAS PROC CALIS and PROC FACTOR, SAS/GRAPH, and SPSS.

RESEARCH EXPERIENCE

Spring 2007 to present, 1997 to 2004 — Rutgers University, Dept. of Psychology
Piscataway, NJ
Graduate Assistant, *Electrophysiological Study of the Basal Ganglia*
- Conducted electrophysiological recordings in a rodent behavioral pharmacology paradigm.
- Used SAS, SPSS, and HLM to develop and implement various univariate and multivariate data analytic strategies for complex multilevel neural and behavioral data.

Fall 2009 to present , Fall 2002 to Summer 2005	Rutgers University, Center of Alcohol Studies Piscataway, NJ **Research Assistant,** *Study of Neuropsychological Factors in Alcoholism Treatment* • Analyzed data from clinical trials of alcoholism treatment using multi-group structural equation models. • Investigated the effects of neuropsychological impairment on successful alcoholism treatment outcome (using mediation and moderation models) and on psychopathology assessment (using differential item functioning). • Implemented complex statistical models using Mplus, SAS, and SPSS.
Fall 2005 to Spring 2008	Rutgers University, Office of Institutional Research and Planning New Brunswick, NJ **Research Assistant,** *Management and Analysis of Institutional Data* • Conducted studies on the institutional characteristics of Rutgers University and other comparable academic institutions. • Analyzed and managed institutional data using SAS, SPSS, and Crystal Reports.
Winter / Spring 1996	Rutgers University, Dept. of Psychology Piscataway, NJ **Research Assistant,** *AIDS Prevention Study* • Programmed PARADOX for data entry and analysis. • Entered qualitative interview data into PARADOX. • Trained undergraduates in operation of PARADOX.

PUBLISHED REFEREED ARTICLES

Pawlak, A.P., Tang, C., Pederson, C., Wolske, M.B., & West, M.O. (2010). Acute effects of cocaine on movement-related firing of dorsolateral striatal neurons depend on baseline firing rate and dose. *Journal of Pharmacology and Experimental Therapeutics,* 332: 667–683.

Root, D.H., Fabbricatore, A.T., Barker, D.J., Ma, S., Pawlak, A.P., & West, M.O. (2009). Evidence for habitual and goal-directed behavior following devaluation of cocaine: a multifaceted interpretation of relapse. *PLoS One,* 4**:** e7170

Tang, C., Mittler, T., Duke, D.C., Zhu, Y., Pawlak, A.P., & West M.O. (2008). Dose- and rate-dependent effects of cocaine on striatal firing related to licking. *Journal of Pharmacology and Experimental Therapeutics,* 324: 701-713.

Tang, C., Pawlak, A.P., Prokopenko, V., & West, M.O. (2007). Changes in activity of the striatum during formation of a motor habit. *European Journal of Neuroscience,* 25: 1212-1227.

Bates, M.E., Pawlak, A.P., Tonigan, J.S., & Buckman, J.F. (2006). Cognitive impairment influences drinking outcome by altering therapeutic mechanisms of change. *Psychology of Addictive Behaviors,* 20(3): 241-253.

Prokopenko, V.F., Pawlak, A.P., & West, M.O. (2004). Fluctuations in somatosensory responsiveness and baseline firing rates of neurons in the lateral striatum of freely moving rats: Effects of intranigral apomorphine. *Neuroscience*, 125: 1077–1082.

Ghitza, U.E., Fabbricatore, A.T., Prokopenko, V., Pawlak, A.P., & West, M.O. (2003). Persistent cue-evoked activity of accumbens neurons after prolonged abstinence from self-administered cocaine. *Journal of Neuroscience*, 23(19): 7239-7245.