DIFFERENTIAL ITEM FUNCTIONING IN THE K-SADS USING DIAGNOSIS OF MAJOR DEPRESSIVE DISORDER AS A GROUPING VARIABLE

By

Corinne Elizabeth Perkins

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

May, 2013

Nashville, Tennessee

Approved:

David A. Cole, Ph.D.

Sun-Joo Cho, Ph.D.

Andrew Tomarken, Ph.D.

Bahr Weiss, Ph.D.

To my brilliant and adorably strong willed daughter, Aurora Rose

and

To my beloved husband, Perky, who is infinitely supportive. Chi!

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisor, Professor David Cole, for his infectious enthusiasm and the tremendous knowledge he shares with his students. His extraordinary patience and sense of humor made frustrating obstacles fun learning experiences. I could not have navigated my way through the graduate school milestones without him.

I would also like to thank my committee members: Dr. Sun-Joo Cho, Professor Andrew Tomarken, and Professor Bahr Weiss for their encouragement, insightful comments, and hard questions.

On a more personal note, I would like to thank my family for the encouragement and support through this process. I would particularly like to thank my Mom who has always been my favorite confidant. Her support is unconditional and her insight is invaluable.

Thank you to all my Vanderbilt friends who are now a part of my family. Particularly the Quant lunch gang: Dr. Michele Tomlinson, Dr. Laura Williams, Andriy Koval, Stijn Smeets, and least of all, Michael Nelson. You all shared your quantitative wisdom with me every week and offered equally horrifying graduate school stories to let me know that I was not alone. I love you all; and I cannot wait to see all the amazing things you accomplish in the future.

Last but certainly not least, I am truly indebted to my husband, Perky, and my daughter, Aurora. You both inspire me to do better, be better, and to reach my goals. You make me laugh and smile at times when it seems impossible. You are the most important things in my life and I owe my successes to your unbelievable support and love. Chi! Thank you all.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS AND SYMBOLS

α	Discrimination / Slope
β	Location / Threshold
CI	Confidence Interval
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IRF	Item Response Function
IRT	Item Response Theory
K-SADS	Kiddie Schedule for Affective Disorders and Schizophrenia
MDD	Major Depressive Disorder
PL	Parameter Logistic
TIF	Test Information Function

CHAPTER I

Introduction

Measures of psychopathology that were originally developed to facilitate a categorical diagnosis are often also used to assess severity, but this process assumes that the psychometric properties of the measure are valid and psychometrically invariant across populations with and without the disorder. If such measurement invariance does not exist, then cross-group comparisons can be highly misleading. True group differences can be missed. Spurious group differences can be found. In psychopathology research, the discovery of a significant psychopathology correlate could be the result of measurement non-invariance across the depressed and nondepressed individuals contained in the sample. In clinical treatment research, measurement non-invariance could either mask or exaggerate true treatment effects, potentially generating misleading conclusions. The current study examines whether or not the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children (K-SADS; a measure of depression in children) is psychometrically invariant across individuals with and without major depression.

Differential Item Functioning (DIF)

Within the Item Response Theory (IRT) literature, a family of procedures called Differential Item Functioning (DIF) can be used to test measurement invariance across groups. DIF analysis makes the distinction between true group differences on the latent dimension of interest (called *impact* in the IRT literature) and artifactual group differences that are due to measurement invariance problems. Within a 2-parameter logistic model DIF, measurement non-invariance can occur in two forms: DIF in item *location* and *discrimination*. The item location, also commonly referred to as an *item difficulty* in IRT applications for cognitive tests, describes the point on the latent dimension where the probability of people endorsing the item is .5 (see Figure 1). For the current study, depression is the latent dimension, and item endorsement reflects the presence of a symptom. Given this conceptualization, a larger item location indicates that people who endorse that symptom likely have greater depression severity. Item discrimination refers to the slope of the item characteristic curve (ICC) at the item location, which represents the strength of relation between the item (or symptom) and the underlying latent variable (depression; see Figure 1). Higher discrimination indicates a steeper ICC slope and implies that the item does well discriminating between people below and above the location of the item.

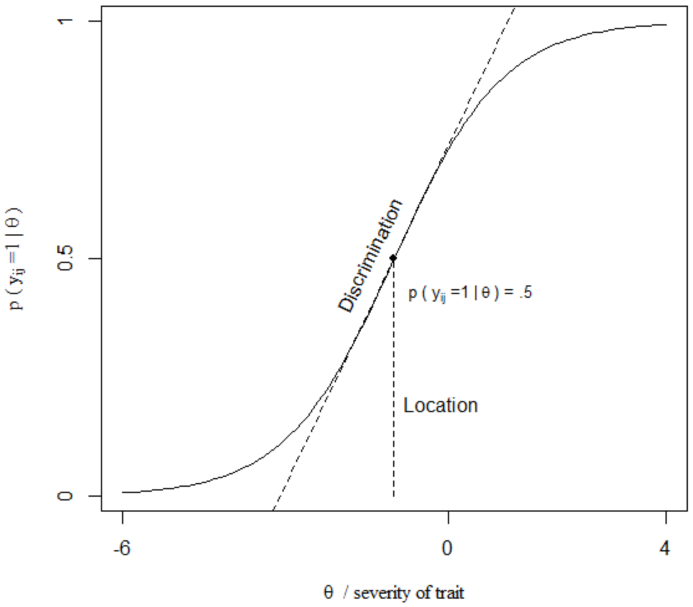

Figure 1. Item location and discrimination labeled on an item characteristic curve. The y-axis, $p(y_{ij} =1|\theta)$, denotes the probability that an individual will endorse the specific item the graph is describing given the individual's depression severity.

In order for clinical researchers to focus on impact, the items that comprise clinical measures must not show DIF in discrimination or location when controlling for the level of latent dimension. DIF in location or discrimination likely indicates that the two IRT assumptions (local independence and unidimensionality in the latent space being measured) have been violated. The presence of DIF indicates that some source of variance other than the latent construct intended to be measured has influenced item responses. Thus, if a measure has items that function differentially across groups, then researchers should be concerned that the measure is assessing different latent constructs for different populations. Furthermore, when items exhibit DIF, the IRT assumptions have been violated, and further analysis can lead to unreliable IRT parameter estimates and erroneous statistical outcomes. These problems are particularly important when attempting to identify group differences. Researchers who use measures with DIF should be thoughtful about both the possible causes and consequences of DIF before utilizing such a measure in their study.

A hypothetical example of location DIF involves the use of self-reported crying as an indicator of depression in boys and girls. Specifically, girls may endorse an item about crying when they have a relatively low level of depression severity; however, due to social pressure boys may require a higher depression severity before they endorse a crying item. The fact that a construct other than depression, in this case social pressure, influenced boys' responses indicates that DIF swayed boys to respond differentially than girls. An ICC for girls and boys the curves will look identical but the boy's curve would be further to the right of the depression dimension (see the left side of Figure 2).

A hypothetical example of discrimination DIF involves a particular item on the quantitative section of a scholastic aptitude test. Imagine that the item included a picture of a seesaw, and the purpose of the question was to assess students' understanding of equality when differently shaped and weighted items were put on both sides of the seesaw. This item might be a less valid measure of quantitative aptitude for urban children compared to suburban children because urban children have less access to playgrounds and less firsthand experience with seesaws. If you drew an ICC for suburban and urban children the slope of the item for suburban children would likely be steeper than the slope for urban children (see the right side of Figure 2). Moreover, the slope differences between the two groups depicts that when an urban and suburban child have identical math abilities that are below the location of the item, the urban child is more likely to answer the question correctly; however, when suburban and urban children have identical math abilities that are above the location of the item, suburban children are more likely to answer correctly. This example illustrates that when there is DIF discrimination, individuals' group membership and their location on the continuum compared to the item location determines if they are more or less likely to answer an item correctly compared to the other group.

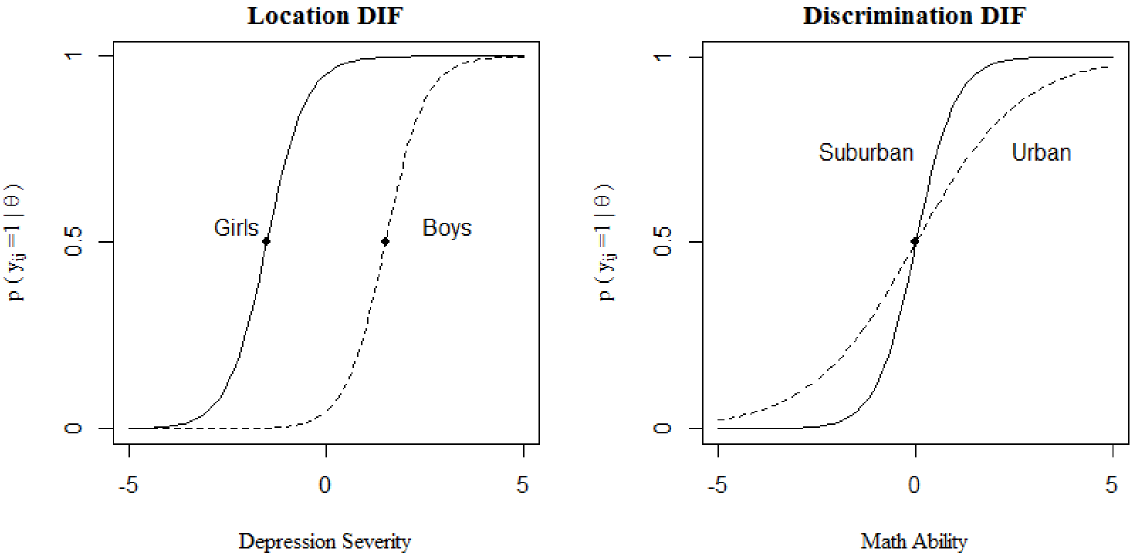

Figure 2. Hypothetical item characteristic curves showing examples of location and discrimination differential item functioning. The y-axis, $p(y_{ij} =1|\theta)$, denotes the probability that an individual will endorse the specific item the graph is describing given the individual's depression severity.

Clinical Diagnosis and DIF

Traditionally DIF analysis uses grouping variables that are distinct from the construct being measured. For instance, in the first example, the concern was gender DIF on a depression item. In the second example the concern was urban/suburban DIF on a math aptitude item. The current study is unique because the grouping variable is major depressive disorder (MDD) and the latent continuum is depression. Until now, no study has conducted a DIF analysis using a grouping variable that is closely related to the underlying latent variable of interest.

Real consequences exist if groups with and without diagnosis respond differentially to items that are meant to assess the severity of the disorder. Those doing psychopathology and clinical research should be particularly concerned about DIF in their measures because their studies usually involve individuals moving from a one group to another. Three common research designs in these fields are (1) comparing those with and without the diagnosis, (2) implementing a treatment to test whether diagnosed individuals no longer meet diagnosis criteria, and (3) following individuals over time to identify naturally occurring variables that may contribute to the onset, remission, or relapse of a disorder. All three designs involve the assumption that the measure used to assess the disorder is psychometrically invariant across people with and without the disorder. Measurement invariance likely occurs because despite identical questions and measurement presentation, the measure assesses a different latent construct for people with diagnosis than for individuals who do not meet criteria. Measurement invariance is an essential assumption of clinical research and researchers should analyze their measurements to ensure that DIF is not contributing to misleading results.

The Current Study

The overarching goal of the current study was to determine whether there is evidence of location or discrimination DIF or both in any of the depression symptoms on the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children (K-SADS) between nonMDD and MDD children and adolescents. Unfortunately, this DIF analysis proved to be impossible due to scaling issues within the MDD group with an IRT approach. Consequently, a creative solution was implemented whereby the nonMDD group was compared to the total sample (comprised of both the nonMDD and MDD groups). Results from a DIF analysis comparing the nonMDD group to the total sample would imply that the MDD group is responsible for deviant item parameters, as the MDD data are all that differentiate between the nonMDD group and total sample. Note that the culpability of the MDD group can only be inferred from the DIF analysis and that this analysis is not equivalent to a DIF analysis that directly compares two mutually exclusive groups, nonMDD and MDD groups.

The current study is important to researchers because the K-SADS, like most diagnostic measures, has historically been assumed to measure the same construct in populations where diagnosis is present and absent. If results indicate no DIF, then the nonMDD group can be thought to exist on the same unidimensional depression severity continuum as the total sample (which contains the MDD group). Furthermore, it offers support for the hypothesis that the K-SADS is measuring the same construct for both nonMDD and MDD populations. More importantly, lack of DIF provides evidence of measurement invariance, suggesting that the K-SADS is less likely to generate misleading conclusions by masking or exaggerating true group differences.

Finally, the results of this study also have important implications for clinicians. Because the K-SADS is a valid and reliable semi-structured interview with parents and children, it is as close as we have to a gold standard method for assessing depression in children and adolescents. Each item of the K-SADS represents a symptom of the underlying disorder. Consequently, the discovery of DIF between the nonMDD group and total sample would not only reveal serious measurement issues but would begin to suggest that the underlying latent variable structure of depression itself differs depending on whether or not an individual has the disorder.

CHAPTER II

Methods

Participants / Data Selection

IRT analysis, including IRT DIF analysis, requires a large sample size to obtain reliable parameter estimates. Hence, the current study aggregated 14 data sets from numerous sites around the world to create a large and diverse sample (see Cole et al., 2011 for specific details regarding this process including originating data set sources). Inclusion in the current study meant participants must have: (a) had symptom-level information gathered through a K-SADS interview from both the child and the parent; (b) been between the ages of 4 and 19 years old when given the K-SADS; and (c) completed the K-SADS prior to any treatment. If participants had multiple interviews that met these criteria, we used the data from the most recent interview. Ultimately, the DIF analysis for this study used data from 3,386 participants.

Measurement

The K-SADS is a clinical interview given to children and their parents to assess a wide variety of psychopathology in children. Variation in the implementation and presentation of the K-SADS exists because instructions for the K-SADS encourage clinicians to prompt and rephrase questions in a way that individual participants can understand; furthermore, many different versions of the K-SADS exist. Despite the flexibility, there remains great similarity across implementation and versions. In fact, Ambrosini's (2000) review of the development and psychometric features of the K-SADS found good interrater reliability and validity to measure and diagnose depression for different versions of the measure. The current study uses only the part of the K-SADS that assesses symptoms of MDD from aggregated data from multiple versions of the K-SADS, including: the K-SADS-Present and Lifetime Version (K-SADS-PL;

Kaufman et al., 1997), K-SADS-PL Version 1.0 (Kaufman, Birmaher, Brent, RAO, & Ryan, 1996), K-SADS-Epidemiological Version (K-SADS-E; Orvaschel, 1994), Washington University in St. Lois K-SADS (WASH-U-K-SADS; Geller et al., 2001), and K-SADS-Version IV-Revised (K-SADS-IV-R; P. Ambrosini & Dixon, 1996).

Variables

Responses to the K-SADS were organized into nine, dichotomous depression symptoms using the basic structure of the interview. For the current study, we refer to the variables as "symptoms" rather than "items." Items are generally regarded as a single question, but in the K-SADS interview, multiple questions are typically asked before he clinician records the presence or absence of a symptom. For this reason, the term "symptom" seems more appropriate than "item." Six of the nine symptoms were *compound* symptoms, meaning that these symptoms are regarded as present if any of several *component* symptoms is evident. The nine symptoms, with example interview prompts and a list of compound breakdowns are as follows:

1. *Depression/irritability.* (a) Depressed mood: Have you ever felt sad, blue, down, or empty? Did you feel like crying? Did you have a bad feeling all of the time that you couldn't get rid of? (b) Irritability: Was there ever a time when you got annoyed, irritated, or cranky at little things? Did you ever have a time when you lost your temper a lot?
2. *Pervasive Anhedonia* (lack of interest, apathy, low motivation, or boredom). Has there ever been a time when you were bored a lot of the time? Did you look forward to doing the things you used to enjoy? Did you have to push yourself to do your favorite activities? Did they interest you?
3. *Weight/appetite disturbance.* (a) Appetite loss: How is your appetite? Do you feel hungry often? Do you leave food on your plate? Do you sometimes have to force yourself to eat? (b) Weight loss: Have you lost any weight since you started feeling sad? Do you find your clothes looser now? (c) Appetite gain: Have you been eating more than before? Is it

like you feel hungry all the time? (d) Weight gain: Have you gained any weight since you started feeling sad? Have you had to buy new clothes because the old ones did not fit any longer?

4. *Sleep disturbance.* (a) Insomnia: Do you have trouble sleeping? How long does it take you to fall asleep? Do you wake up in the middle of the night? Do you wake up earlier than you have to? (b) Hypersomnia: Are you sleeping longer than normal? Do you go back to sleep after you wake up in the morning?
5. *Psychomotor disturbance.* (a) Agitation: Since you've felt sad, are there times when you can't sit still, or you have to keep moving and can't stop? Do people tell you to not talk so much? (b) Retardation: Since you started feeling sad, have you noticed that you can't move as fast as before? Has your speech slowed down? Have you felt like you are moving in slow motion?
6. *Fatigue* (lack of energy and tiredness). Have you been feeling tired? Do you take naps because you feel tired? Do you have to rest? Do your limbs feel heavy? Is it very hard to get going?...to move your legs?
7. *Self-perceptions.* (a) Worthlessness: How do you feel about yourself? Do you like yourself? Do you ever think of yourself as pretty or ugly? Do you think you are bright or stupid? (b) Excessive or inappropriate guilt: Do you feel guilty about things you have not done? Or are actually not your fault? Do you feel you cause bad things to happen? Do you think you should be punished for this?
8. *Cognitive disturbance.* (a) Concentration, attention, slowed thinking: Sometimes children have a lot of trouble concentrating, like [list examples]. Have you been having this kind of trouble? Is your thinking slowed down? When you try to concentrate on something, does your mind drift off to other thoughts? Can you pay attention in school? (b) Indecision: When you were feeling sad, was it hard for you to make decisions?

9. *Suicide.* Sometimes children who get upset or feel bad, wish they were dead or feel they'd be better off dead. Have you ever had these types of thoughts? Sometimes children who get upset or feel bad think about dying or even killing themselves. Have you ever had such thoughts? How would you do it? Do you have a plan?

Missing Data and Outliers

Sites and interviewers differed on whether they used the "skip-out" procedures described in the K-SADS directions. Some of the sites collected data on all of the depression symptoms, no matter the participants' response patterns; whereas other sites asked screening questions, which could lead the interviewer to skip certain follow-up questions. Two skip-out procedures were used by investigators. Some asked participants two screening symptoms: Depression/Irritability and Pervasive Anhedonia. Others asked three screening symptoms (the two mentioned above and an additional question about Suicide). Skip-out procedures created missing data for about 22.5% of the cases. Besides the missing data due to the skip-out procedure there were also random missing data for 5% of cases; however, analysis showed no psychometric differences between the group with random missing data and those with complete data. When fitting the IRT models, we utilized all the acquired responses, rather than eliminating individuals with missing data, by using full information maximum likelihood estimation (Bock & Aitkin, 1981).

A total of 3,403 participants met the three criteria mentioned in the Participants / Data Selection section; however, further investigation identified seventeen outliers. Outliers fell into two categories: either the participants endorsed symptoms in a highly peculiar way or the skip-out protocol was breeched. An example of a peculiar response pattern was when Suicide was the only depressive symptom that a participant endorsed. A skip-out protocol breech was evident when a participant endorsed one of the screening symptoms but was not asked about any more depression symptoms. The 17 cases that included such outliers or patterns were dropped from the study. The final sample included 3,386 participants.

CHAPTER III

Results

Descriptive Statistics

The data used for this study consisted of 1,140 participants who met the DSM-IV criteria for Major Depressive Disorder (MDD) and 2,246 (66.3%) participants who did not (American Psychiatric Association, 2000). These groups shall be referred to as the "MDD group" and "nonMDD group." The "total sample" consisted of 1,712 males and 1,671 females (50.6% and 49.4% respectively, with missing gender data for three participants). The sample was ethnically diverse with 2,226 (65.8%) whites, 800 (23.6%) African Americans, 126 (3.72%) Hispanics, and 214 (6.33%) participants of an ethnicity not specified here (20 participants were missing ethnicity data). As mentioned earlier, inclusion in the study meant that participants were between the ages of 4 and 19 (M = 12.38 years, SD = 2.98). Further partitioning of demographic data by the participants' diagnostic status is in Table 1.

Descriptive Statistics About Item Parameter Estimates

Two issues prohibited the use of IRT models when analyzing the data from the MDD group. The first issue was a lack of variability in response patterns due to the diagnostic criteria put forth in the DSM-IV (American Psychiatric Association, 2000). The second issue was that the data had significant scalability issues (see Appendix A for more details about both of these issues). Neither of the two afore mentioned issues were present for the total sample or the nonMDD group; thus IRT models could be used for analysis.

The R package "difR" was used to fit a 2-Paramter Logistic (2-PL) IRT model using the data from the total sample (see Appendix B for the R code used to for the results section analyses). Figure 3 shows the test information function (TIF) with an overlaid stacked-histogram

Table 1

Frequencies of Demographic Variables of Participants by Diagnostic Classification

	Diagnostic Classification	
Variable	nonMDD[a] (*n* = 2,246)	MDD[b] (*n* = 1,140)
Gender		
Female	1024	647
Male	1219	493
Race		
White	1391	835
African American	643	157
Hispanic	61	65
Other	143	71
Age		
4	1	0
5	39	7
6	69	11
7	105	6
8	140	18
9	219	34
10	192	47
11	248	37
12	364	108
13	204	145
14	172	196
15	258	211
16	164	182
17	45	118
18	7	12
19	0	1

Note. MDD = major depressive disorder.
[a]The number participants missing data for gender, race, and age was 3, 8, and 19 respectively.
[b]The number of participants missing data for race and age was 12 and 7 respectively.

of the total sample separated into nonMDD and MDD groups. The x-axis represents depression severity (also referred to as θ), as severity increases from left or right. The left y-axis denotes the frequency of participants in the histogram and the right y-axis represents the amount of test information. Notice that the majority of nonMDD participants are represented in the very tall bar on the far left of the depression severity continuum; this bar consists of the participants who denied having any of the nine symptoms. People with zero symptoms accounted for 70.9% of the nonMDD group and 47% of the total sample. Figure 3 reveals that there is relatively little overlap between the nonMDD and MDD groups. Figure 3 also shows that the majority of test information is distributed over only a small portion of nonMDD participants. Therefore, the symptoms on the K-SADS offered very little information (or large standard error of the depression severity) about the majority of the people in the nonMDD group.

On the positive side, the most information gathered from the K-SADS, as indicated by the TIF, was fortuitously located where the nonMDD and MDD groups overlap. Put another way, the measure gave the most precise estimates of depression severity and had the smallest standard error in the narrow range where nonMDD and MDD children existed together. Precision in this range is especially important for the diagnostic purpose of distinguishing between individuals who are just above diagnostic threshold from those who are just below threshold for MDD. Therefore, the location of the TIF peak supports the use of the K-SADS in clinical and research settings where rigorous distinctions need to be made between children who nearly meet criteria for MDD.

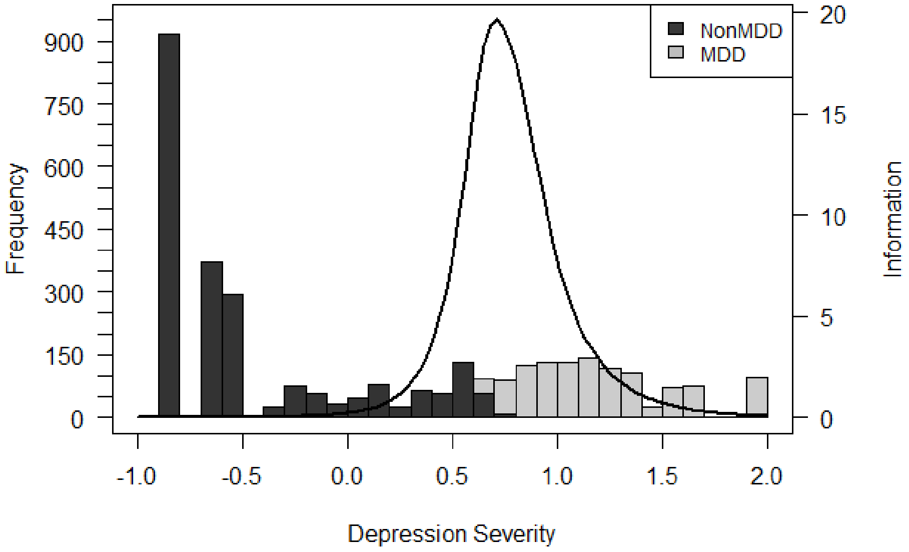

Figure 3. Stacked histogram of the nonMDD and MDD groups with an overlaid test information function estimated using the 2-parameter logistic item response model using the total sample data. MDD=major depressive disorder.

A second 2-PL IRT model was run using the data from the nonMDD participants only (item parameter estimates obtained from the “difR” R package). The estimated item parameters, discrimination (α) and location (β), from the total sample and the nonMDD group are presented in Table 2 (results were equated for comparison). Figure 4 shows the 95% confidence intervals (CI) constructed around the item parameters shown in Table 2. The horizontal axis in Figure 4 represents the range of the CI for discrimination and location respectively on the latent dimension of depression severity, whereas the vertical axis represents the nine symptoms. The CI of the total sample is shown using solid lines and the nonMDD group is represented by dashed lines. For each symptom, the less the CIs for the total sample and nonMDD group overlap, the more likely the symptom will have DIF.

Table 2

Item Parameters from a 2-PL IRT Model that Contained Nine K-SADS Symptoms

	Total Sample (*N* = 3,386)				nonMDD Group (*n* = 2,246)			
	Discrimination (α)		Location (β)		Discrimination[a] (α)		Location (β)	
Symptom	Est.	(S.E.)	Est.	(S.E.)	Est.	(S.E.)	Est.	(S.E.)
Depression/Irritability	5.71	(0.28)	0.25	(0.11)	4.29	(0.40)	0.31	(0.09)
Pervasive Anhedonia	4.91	(0.32)	0.60	(0.20)	4.07	(0.53)	0.73	(0.15)
Weight/Appetite Disturbance	2.01	(0.11)	0.88	(0.09)	2.55	(0.34)	0.83	(0.08)
Sleep Disturbance	3.01	(0.15)	0.48	(0.10)	2.95	(0.32)	0.54	(0.08)
Psychomotor Disturbance	2.33	(0.13)	0.72	(0.09)	3.83	(0.44)	0.51	(0.12)
Fatigue	3.81	(0.21)	0.52	(0.14)	4.20	(0.50)	0.53	(0.14)
Self-perceptions	2.53	(0.12)	0.47	(0.08)	2.76	(0.29)	0.50	(0.07)
Cognitive Disturbance	3.43	(0.17)	0.41	(0.11)	4.27	(0.45)	0.36	(0.11)
Suicide	1.77	(0.10)	1.33	(0.09)	1.95	(0.35)	1.37	(0.08)

Note. MDD = major depressive disorder. Est. = estimate. S.E. = standard error. Estimates across total sample and nonMDD group are on the same scale.

[a]The nonMDD discrimination standard errors are quite large due to the truncated range of depression severities of the nonMDD group and the resulting lack of overlap between the nonMDD participants and the TIF (see Figure 3).

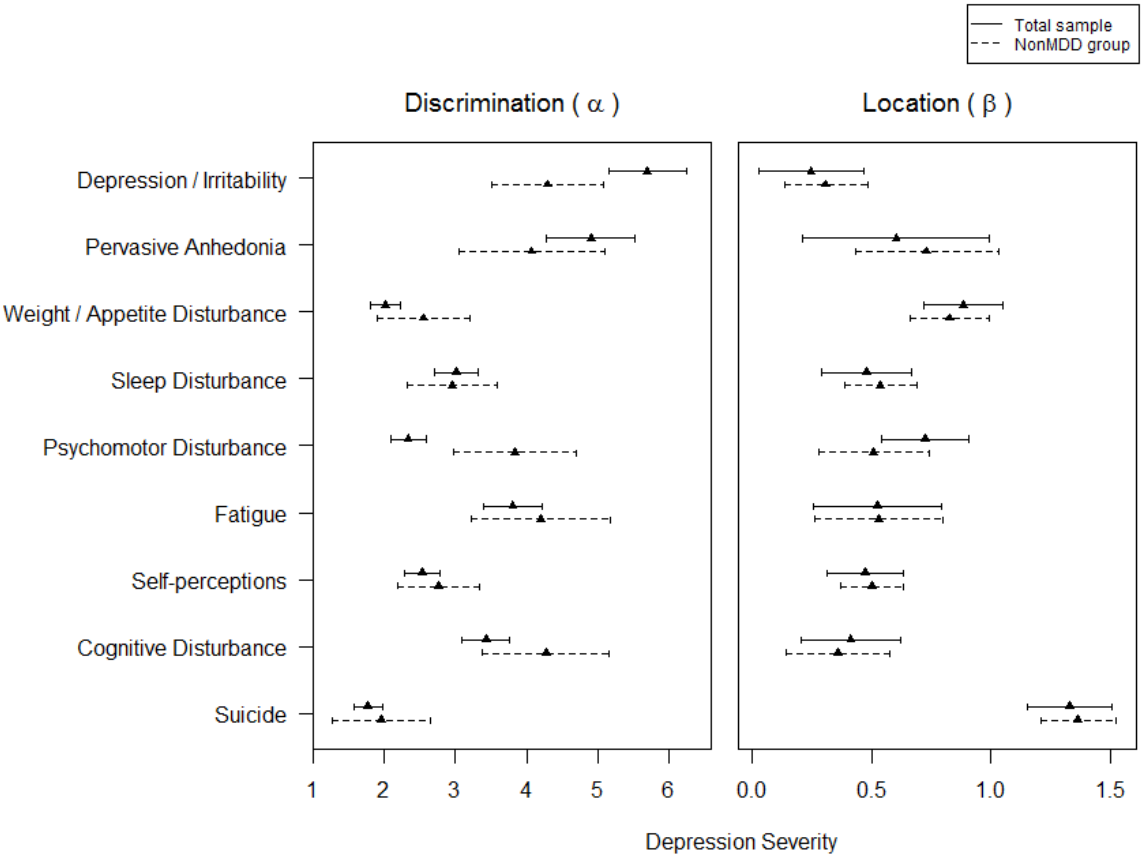

Figure 4. Confidence intervals around item parameters for nine K-SADS symptoms. MDD=major depressive disorder.

Similarities and differences between the item parameter estimates from the total sample and the nonMDD group can also be seen when both groups' item characteristic curves (ICCs) are presented on the same graph. The less similar the ICCs, the more likely the symptom is to have DIF. Figure 5 shows nine graphs, each graph represents a symptom. Within each graph there are two ICCs, one estimated from the 2-PL IRT model that used data collected from the total sample and one estimated from the model using only the nonMDD participants. The x-axis is latent depression (also called θ, severity increases left to right) and the y-axis is the probability of a participant endorsing the symptom at a given level of depression severity. Each ICC was restricted to the estimated range of depression severities in each group (nonMDD group: -0.625 to .477 and total sample: -0.848 to 1.966).

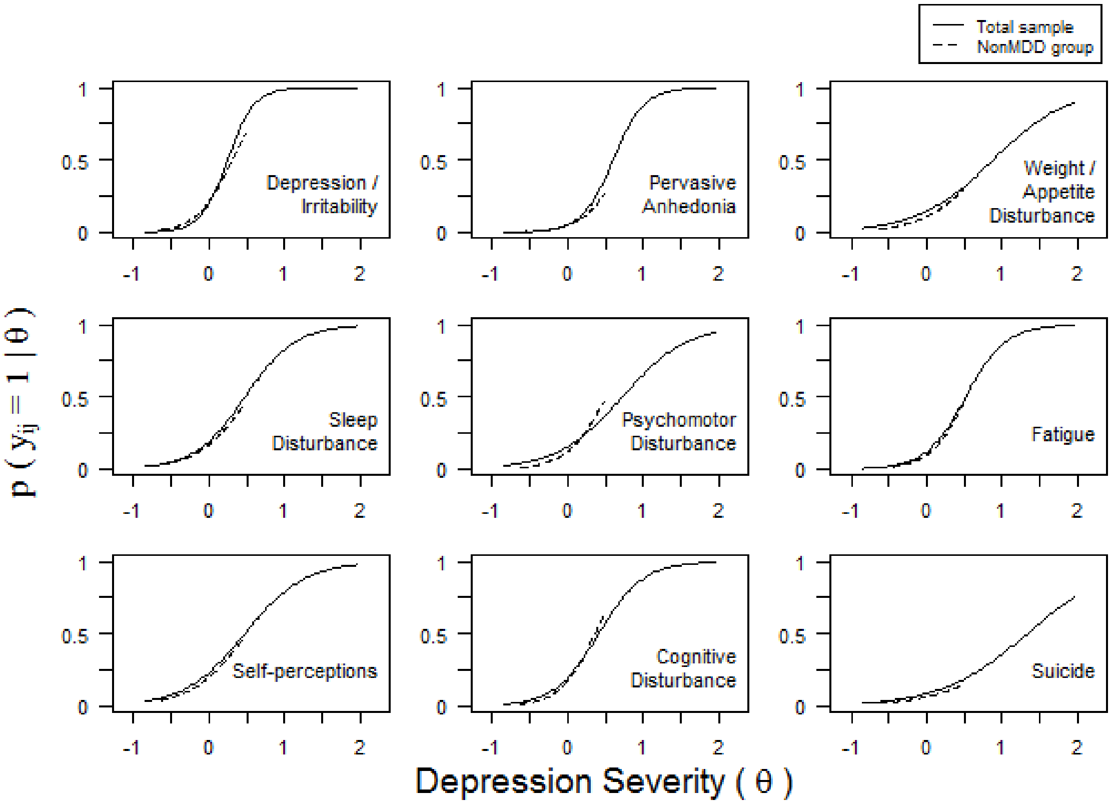

Figure 5. Item characteristics curves for nine K-SADS symptoms, modeling data from the total sample and nonMDD group. The y-axis, p(y_{ij} =1|θ), denotes the probability that an individual will endorse the specific symptom the graph is describing given the individual's depression severity. MDD = major depressive disorder.

Differential Item Functioning

Descriptive statistics, particularly those depicted in Figures 4 and 5, offer visual clues as to which symptom(s) might have DIF. For example, Figures 4 and 5 show the Psychomotor Agitation symptom as having the most obvious indications of DIF, by having blatantly non-overlapping ICCs and discrimination CIs. These descriptive statistic techniques alone are not persuasive enough to make a claim of DIF; researchers must also use concrete DIF detection statistics to make credible DIF claims.

There are three approaches of IRT DIF detection statistics commonly used: (a) comparing the model fit for each group (Thissen, Steinberg, & Wainer, 1988); (b) calculating the area between the ICCs (Raju, 1988); and (c) assessing the difference between item parameter estimates calculated for two groups (Lord, 1977, 1980). These three statistics quantify DIF in very different ways, leading researchers to use more than one statistic to analyze their data. We opted not to use the model comparison approach for this study, due to the nature of our groups. Specifically, when one group is a subset of the other group (e.g., the nonMDD group is part of the total sample), the model comparison approach will always indicate that the larger group fits better. The remaining two DIF statistics, Raju's *Z* (*H*, unsigned area) and Lord's χ^2, were used for this analysis. Raju's *Z(H)* is a DIF detection statistic that calculates the area between the ICCs of the groups of interest (refer to Figure 5). Raju suggested that researchers with sample sizes greater than 500 participants use a critical *Z* score value of ±3 (alpha =.0027) to determine statistical significance, so as not to overpower the statistic (Raju, 1990). Lord's χ^2 acts like a multivariate *t*-test that calculates the difference between item parameters across groups. Although Lord did not make any suggestions about changing the alpha level for larger samples, the same stringent α was used for both statistics in an effort to remain consistent (alpha =.0027, critical χ^2=11.8). Unfortunately, both of these statistics assume that the groups are mutually exclusive and independent; thus they do not incorporate the appropriate covariance terms (between $\hat{b}_1$ and $\hat{a}_2$, $\hat{b}_1$ and $\hat{b}_2$, $\hat{a}_1$ and $\hat{b}_2$, and $\hat{a}_1$ and $\hat{a}_2$) for the current application to dependent groups. Due to the nature of this dependence, the covariance between groups would likely be positive; thereby inflating the *Z* and χ^2 statistics. Thus the results of these statistics must be used with caution, even when using the extremely stringent alpha level of .0027. Furthermore, because these statistics quantify DIF using different methods, they sometimes yield discrepant results. In fact, Cohen and Kim (1993) found in a simulation study that Raju's *Z* (both exact signed and unsigned area) was less effective at detecting DIF than Lord's χ^2. Generally, items that demonstrate DIF across multiple detection statistics are especially noteworthy.

The results of the Raju's *Z(H)* and Lord's χ^2 DIF analyses are shown in Table 3. Notice that Raju's *Z(H)* did not identify any of the symptoms as having DIF between the total sample and the nonMDD group; however, Lord's χ^2 identified two symptoms, Depression/Irritability and Psychomotor Disturbance, as having significant DIF. These results indicate that, although the area between the ICCs is not significantly different from zero, the cumulative difference between item parameters is likely something greater than zero. Based on Cohen and Kim's findings that Raju is less effective at detecting DIF, it is not surprising that Lord's χ^2 would identify DIF for symptoms that Raju did not (1993). Despite the fact that the DIF detection statistics used for this analysis were unable to indicate which item parameter(s) contributed to the significant results, looking again at Table 2 and Figure 4, it appears as if the discrimination parameter for both symptoms was the strongest contributor to the significant DIF (as the CIs for this parameter fail to overlap for either symptom). For the nonMDD group, the Depression/Irritability symptom was less discriminating than it was for the total sample; whereas the Psychomotor Disturbance symptom was more discriminating for the nonMDD group.

Table 3

Differential Item Functioning Statistics for Nine K-SADS Symptoms

Symptom	Raju's *Z(H)*	Lord's χ^2
Depression/Irritability	1.32	17.04**
Pervasive Anhedonia	0.49	5.95
Weight/Appetite Disturbance	- 1.45	2.01
Sleep Disturbance	0.26	0.17
Psychomotor Disturbance	- 2.47	11.95*
Fatigue	- 0.32	0.55
Self-perceptions	- 0.75	0.35
Cognitive Disturbance	- 0.89	3.44
Suicide	- 0.59	0.43

**p*-value ≤ .0027. ***p*-value ≤ .001.

Interestingly, both of the symptoms that had a significant χ^2s were compound symptoms; that is symptoms that were a function of two or more other symptoms. Therefore, we broke down both of these compound symptoms into their component parts to identify which component symptoms were responsible for the DIF. The Depression/Irritability symptom was split into Depressed Mood and Irritability. The Psychomotor Disturbance symptom was split into

Table 4

Differential Item Functioning Statistics for Four K-SADS Component Symptoms

Symptom	Raju's *Z(H)*	Lord's χ^2
Depressed Mood	- 0.18	0.46
Irritability	- 1.52	15.24**
Psychomotor Agitation	- 3.77**	19.06**
Psychomotor Retardation	- 1.94	6.12

**p*-value ≤ .0027. ***p*-value ≤ .001.

Psychomotor Agitation and Psychomotor Retardation. We repeated the Raju *Z(H)* and Lord χ^2 statistics were utilized again; however, this time the model consisted of the seven symptoms that did not previously show DIF plus the four new component symptoms (11 symptoms total). As none the original seven symptoms were statistically significant, results for only the four component symptoms are shown in Table 4. Results show that the Lord χ^2 , but not Raju's *Z(H)*, identified the Irritability symptom as having DIF. Additionally, both statistics identified Psychomotor Agitation as having DIF.

Results for the Irritability and Psychomotor Agitation component symptoms were mixed. Item parameters (shown in Table 5) and their CIs (shown in Figure 6) show neither symptom to be the obvious contributor to the significant *Z(H)* and χ^2 results. The CIs for both of the

Table 5

Item Parameters for the Four K-SADS Component Symptoms

	Total Sample (*N* = 3,386)				nonMDD Group (*n* = 2,246)			
	Discrimination (α)		Location (β)		Discrimination[a] (α)		Location (β)	
Symptom	Est.	(S.E.)	Est.	(S.E.)	Est.	(S.E.)	Est.	(S.E.)
Depressed Mood	5.08	(0.28)	0.48	(0.16)	4.37	(0.47)	0.45	(0.21)
Irritability	3.26	(0.17)	0.56	(0.10)	3.92	(0.38)	0.38	(0.15)
Psychomotor Agitation	1.75	(0.12)	1.32	(0.10)	3.39	(0.45)	0.75	(0.23)
Psychomotor Retardation	2.13	(0.13)	1.03	(0.11)	2.57	(0.34)	0.87	(0.16)

Note. MDD = major depressive disorder. Est. = estimate. S.E. = standard error. Estimates across total sample and nonMDD group are on the same scale.

[a] The nonMDD discrimination standard errors are quite large due to the truncated range of depression severities of the nonMDD group and the resulting lack of overlap between the nonMDD participants and the TIF (see Figure 3).

Irritability item parameters appear to have similar amounts of overlap and nearly equivalent means between the two groups (when taking CI ranges into consideration). In the case of Psychomotor Agitation the CIs for the discrimination item parameter did not overlap, making item discrimination an obvious DIF contributor. Interestingly, however, the CIs around the location parameters were nearly non-overlapping as well. Due to the fact that neither item parameter was the obvious, single DIF contributor, it seems reasonable to consider both the discrimination and location item parameters for these two symptoms as possible contributors to DIF. Follow-up examination of the results indicated that Irritability and Psychomotor Agitation were easier and more discriminating for the nonMDD group than for the total sample. Figure 7 shows the overlapping ICCs for the four component symptoms. The Psychomotor Agitation ICCs are the most visually discrepant in the entire analysis.

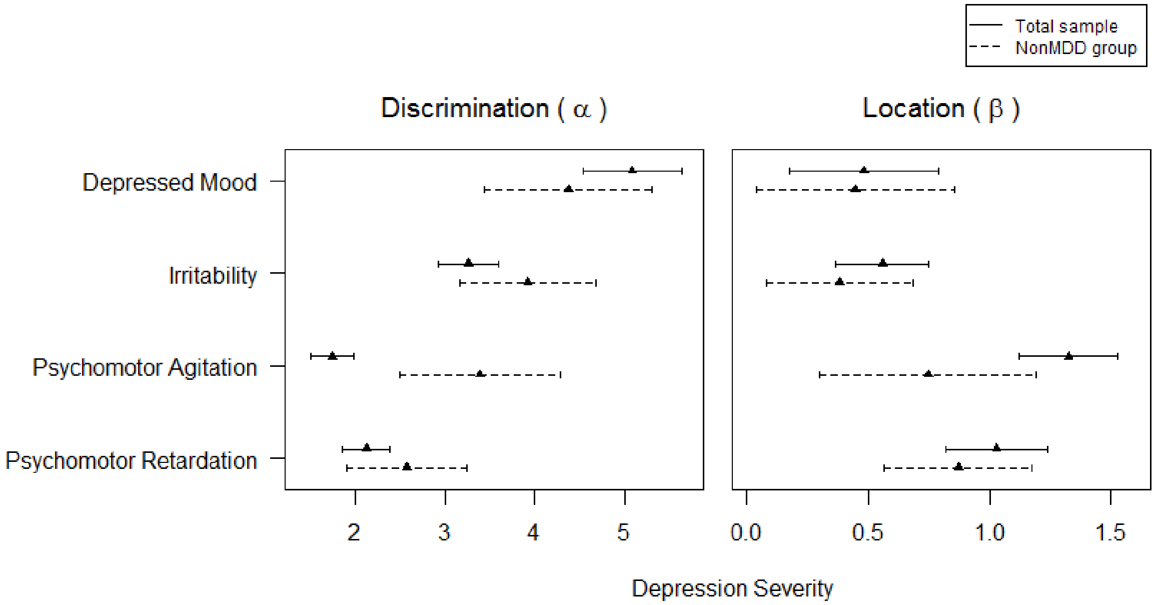

Figure 6. Confidence intervals around item parameter estimates for the four component symptoms that were broken down from component symptoms. MDD=major depressive disorder.

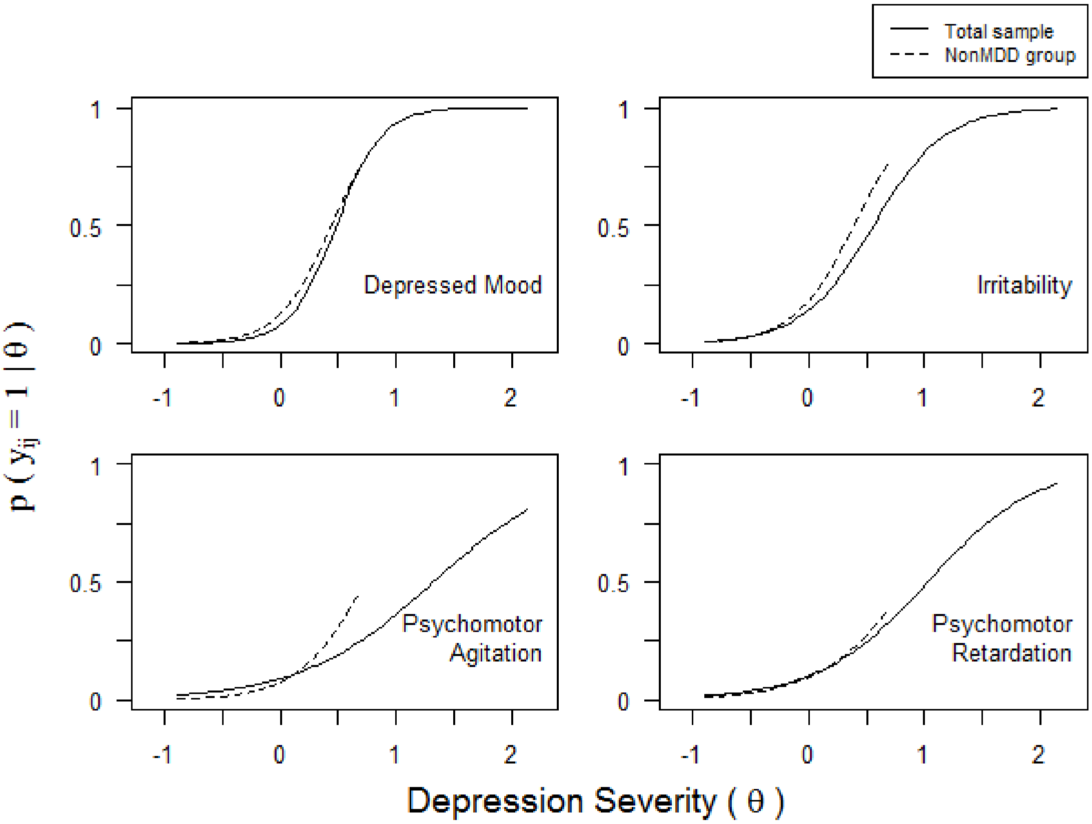

Figure 7. Item characteristics curves for the four component symptoms that were broken down from component symptoms, modeling data from the total sample and nonMDD group. The y-axis, p(y_{ij} =1|θ), denotes the probability that an individual will endorse the specific symptom the graph is describing given the individual's depression severity. MDD = major depressive disorder.

CHAPTER IV

Discussion

The original purpose of this study was to explore the possibility of DIF in the K-SADS across nonMDD and MDD groups. Unfortunately, item parameters could not be estimated in the MDD group. Consequently, we adopted a unique approach to the research question by comparing the nonMDD group to the total sample (i.e., nonMDD and MDD groups combined) rather than the MDD group alone. Because the total sample differs from the nonMDD group only by virtue of the MDD cases, any differences in item parameter estimates between the total sample and nonMDD group can be attributed to the MDD group's response patterns.

Interpreting Results

The major finding in this study was that DIF existed for two symptoms from the K-SADS. Both were compound symptoms: Depression/Irritability and Psychomotor Agitation/ Retardation. Further investigation revealed that Irritability was primarily responsible for the Depression/Irritability DIF and that Psychomotor Agitation was primarily responsible for the Psychomotor Agitation/Retardation DIF. Although Lord's χ^2 identified both of these symptoms as having statistically significant DIF, the Raju *Z(H)* found only Psychomotor Agitation DIF to be significant. Contradictory results from DIF detecting statistics are not uncommon, especially because these particular statistics measure DIF using very different approaches. Furthermore, previous research has shown that Raju's *Z(H)* is less effective at detecting DIF than Lord's χ^2, suggesting that the nonsignificant *Z* should be a relatively small concern (Cohen & Kim, 1993). Lord's χ^2 can be conceptualized as a multivariate *t*-test comparing the item parameters between groups, whereas Raju's *Z(H)* measures the area between the ICCs of the two comparison groups. Both of these DIF detection statistics can be conceptualized as omnibus tests, in that the significance of the statistic does not indicate which item parameter (discrimination or location) is responsible for the DIF. To infer which parameter(s) might have contributed to the significant result, we examined the (non)overlap of the confidence intervals for each item parameter and the ICCs.

For the Irritability component, DIF appeared to be caused by a combination of both item parameters. Overall, the irritability symptom was easier and more discriminating for the nonMDD group than for the total sample. This means that for the range between the intersection of the ICCs and the rightmost point where the nonMDD ICC curve ends (the highest depression severity the nonMDD participants reached; see Figure 7), the nonMDD group was more likely to endorse the symptom than the total population, when depression severity was statistically controlled. Similarly, when depression severity was controlled, the Irritability symptom component was more discriminating for the nonMDD group.

We speculate that the Irritability symptom DIF could be related to depressive attributional style. In both youths and adults, depression is associated with a tendency to attribute negative events to internal causes (Abramson, Seligman, & Teasdale, 1978; Gladstone & Kaslow, 1995). On the other hand, irritability or anger can reflect a tendency to attribute negative events to others or to blame things outside oneself (Dodge & Somberg, 1987; Steinberg & Dodge, 1983). Thus, an individual with depression might be less likely to endorse feelings of irritability. In the current study, the presence of MDD in the total sample (and their absence from the nonMDD group) could be responsible for the greater difficulty of the irritability symptom in the total sample and its greater discrimination in the nonMDD group. That is, for the depressed people in the total sample, a depressive attributional style may make it more difficult to endorse the Irritability symptom, at least to the extent that one’s irritability focuses on others and not oneself.

The current study also found that the Psychomotor Agitation symptom was significantly easier for the nonMDD group to endorse, and more discriminating, compared to the total sample. Again, this means that when depression severity was statistically controlled for the range between the ICC intersections and the most severe nonMDD participants, the nonMDD group was more likely to report psychomotor agitation than was the total population. Furthermore, the symptom was better at discriminating individuals in the nonMDD group, given depression severity. Interestingly, Psychomotor Retardation did not show evidence of DIF between the groups. This indicates that these two symptoms are likely not “two-sides of the same coin,” or in this case, the same latent continuum. Psychomotor retardation was equally likely to be endorsed by both groups when depression severity was statistically controlled. In contrast, the nonMDD group was more likely to endorse feelings of Psychomotor Agitation than the total sample, given depression severity. The disconnect between Psychomotor Agitation and Retardation offers an interesting perspective into the experiences of individuals along latent depression and how depression severity alone may not be able to account for feelings of Psychomotor Agitation.

Perhaps the most surprising aspect of these findings was that, when depression severity was controlled, Irritability and Psychomotor Agitation were easier for the nonMDD group to endorse than the total sample. Given that endorsing these symptoms acts as evidence towards the diagnosis of MDD (American Psychiatric Association, 2000), we would expect that these symptoms would be more difficult for the nonMDD group to endorse compared to the total sample; however, at certain levels of depression severity, the exact opposite was true. The ICCs of Irritability and Psychomotor Agitation, shown in Figure 7, illustrates that for the range between the intersection of the ICCs and the rightmost point of the nonMDD ICC, the nonMDD group has a higher probability of endorsing that symptom than the total sample when controlling for depression severity. This paradoxical finding means that for a range where the nonMDD and MDD participants overlap endorsing the symptoms of Irritability and Psychomotor Agitation is more characteristic of the nonMDD group than the total sample.

Another way to think about the results of this study is that DIF in Irritability and Psychomotor Agitation is evidence of multidimensionality in K-SADS depression scale. Multidimensionality would imply that these two symptoms are assessing more than the unidimensional MDD continuum; at least one dimension other than MDD is being measured by these symptoms. A possible explanation comes from researchers who have argued for the creation of a depression subtype called “agitated depression.” The definition of agitated depression varies slightly in the literature, but two of the most consistent symptoms are irritability and psychomotor agitation (Benazzi, Koukopoulos, & Akiskal, 2004; Benazzi, 2004a, 2004b; Leventhal, Pettit, & Lewinsohn, 2008). Additionally, previous research indicates that agitated depression belongs on the bipolar dimension rather than the MDD dimension (Benazzi et al., 2004). These findings suggest that the K-SADS may have measured the bipolar dimension; however, more research is needed to determine exactly what additional dimension(s) was measured with MDD.

Implications of Results

Despite the Irritability and Psychomotor Agitation symptoms, the K-SADS demonstrated overwhelming unidimensionality among the remaining symptoms. If Irritability and Psychomotor Agitation were removed, strong evidence emerges that the K-SADS meets the requirements for measurement invariance based on the grouping variable of MDD diagnosis. Researchers and clinicians should be concerned about the measurement invariance of the K-SADS when assessing children with and without MDD. If measurement invariance did not exist, then the K-SADS would be quantifying different latent variables for the two groups of children, thereby making it impossible to compare the groups according to the K-SADS results. Thus, findings from this study indicate that without the Irritability and Psychomotor Agitation symptoms, the K-SADS is in fact measuring the same latent depression for both nonMDD and MDD groups (inferred from the total sample).

Psychopathology and clinical researchers often use research designs that (1) compare children with and without the diagnosis, (2) test whether treatment moves people from MDD to nonMDD, and (3) follow individuals over time to identify naturally occurring variables that may correlate with onset and offset of MDD. These methods often assume that individuals will slip in and out of episode; furthermore, these methods assume that the measure used to assess the disorder is psychometrically invariant across people with and without the disorder. If measurement invariance does not exist, results could be misleading. Specifically, treatment effects could be exaggerated or masked because of the inequitable measurement responses. Pending confirmatory research, results from this study could imply that researchers should refrain from collecting (or using) data from the Irritability and Psychomotor Agitation symptoms from the K-SADS to measure depression severity. The remaining symptoms on the K-SADS have shown measurement invariance and are safe to use when comparing nonMDD and MDD children.

Red Flags

As with all research, there are limitations to the current study. One limitation is the unique relation between the K-SADS and the grouping variable. To the knowledge of the author, all previous studies that conducted a DIF analysis used a grouping variable that was not so dependent on the latent dimension that the IRT model estimated. For example, a typical DIF analysis estimating depression severity might use gender, race, or age as a grouping variable. In contrast, the current study used MDD diagnosis as the grouping variable while estimating depression severity based on K-SADS responses. Depression severity (estimated from the K-SADS) and MDD diagnosis are extremely correlated, making this DIF analysis unlike any other one done before it. The statistical consequences of conducting a DIF analysis using a correlated latent dimension and grouping variable are unclear, but it is conceptually obvious that there is a rather cyclical framework at play.

A second limitation of this study is that the groups compared in the DIF analysis was the nonMDD group and total sample. This was due to the unscalability of the K-SADS in the pure MDD group. Although this analysis offered insight into the unidimensionality of the nonMDD group and the total sample, which was used to infer about differences between nonMDD and

MDD groups, it did not allow for the direct comparison of MDD and nonMDD groups, rendering the results somewhat difficult to interpret.

The third and most significant limitation of this study was that the authors were unable to identify any DIF detection statistics that did not assume independence of groups. Therefore, both DIF statistics chosen for this study, the Lord's χ^2 and Raju's *Z(H)*, overestimated the DIF by not accounting for dependence. To compensate for this limitation, and the large sample size, a very stringent alpha of .0027 was used, but it is unknown to what extent this compromise actually decreased Type I error rates; and furthermore, to what extent it increased Type II errors.

Future Directions

At least four directions emerge for future research. The first future direction would be to create a DIF detection statistic that allows for dependence between the groups. The missing covariance terms (between $\hat{b}_1$ and $\hat{a}_2$, $\hat{b}_1$ and $\hat{b}_2$, $\hat{a}_1$ and $\hat{b}_2$, and $\hat{a}_1$ and $\hat{a}_2$) in both Raju's *Z(H)* and Lord's χ^2 is a major limitation in this study. Creating DIF detection statistics with these added covariance terms would allow researchers a chance to replicate the findings in this study. Furthermore, future researchers with dependent groups will benefit from these statistics.

The second future direction would be to use a measure other than the K-SADS. By analyzing only one measure of depression it is not clear whether or not the DIF we found is specific to the K-SADS. If a different measure of depression found the same symptoms to have DIF then the symptoms themselves (not K-SADS items) would be to blame, suggesting that agitation and irritability are likely tapping into something besides major depression.

A third possible future direction would be to explore whether there is a relation between the irritability symptom and attributional style, as proposed earlier. A depressive attributional style could make it difficult for a person with MDD to endorse irritability, especially if that irritability reflects an external attribution of blame. Assessing irritability by multiple methods (some of which are not subject to attributional style) while measuring individual differences in attributional style would be one way to examine this possibility. Finding the root cause of the DIF in the irritability symptom is particularly important because it is one of the screening symptoms.

And lastly, the fourth future direction would be to investigate whether the DIF found in Irritability and Psychomotor Agitation was due to diagnoses other than MDD. One such disorder might be agitated depression; another might be bipolar disorder. All of these future directions would add to the current finding and help minimize the limitations already discussed.

APPENDIX A

Difficulties with the MDD Group

The original goal of the current study was to do a straightforward DIF analysis between nonMDD and MDD groups. Unfortunately, two unexpected problems arose from the data of MDD group that prevented this direct analysis from being accomplished. Firstly, the MDD group did not have enough variation in their response patterns to allow the IRT models to estimate reasonable item parameters. Lack of variation in the MDD group's response patterns was likely due to the artifactual requirements of obtaining a MDD diagnosis. According to the DSM-IV, a person must endorse at least five of the nine depression symptoms identified on the K-SADS, and at least one of those endorsements must come from the Depression/Irritability or Pervasive Anhedonia symptoms (American Psychiatric Association, 2000). Diagnostic requirements eliminated 55.67% of the possible item response patterns for the MDD group. Therefore, whereas the nonMDD participants could theoretically respond with any conceivable response pattern (because diagnosis is ultimately the decision of the clinician), the MDD group had less than half of the response patterns afforded to them.

The second reason the MDD group could not be used in analysis was that they violated a main expectation of IRT, that the total score (raw score) and theta ability prediction (in this case "depression severity") should monotonically increase (Embretson & Hershberger, 1999; Schuur, 2003). Mokken's Loevinger *H* is an item scalability coefficient that allows researchers to identify significant violations of manifest (raw score) monotonicity in the scale. The statistic uses the range and slope of the item response functions as well as the distribution of severities to create the coefficient. When the total sample, the nonMDD group, and the MDD group were analyzed independently using Loevinger *H*, the MDD group showed significant violations of monotonicity for every symptom, except the Psychomotor item. Table A1 presents the *H* value for each symptom. In general, the *H* values were very small for the MDD group but much larger for the total sample and non-MDD group. Figure A1 shows the item response functions for the nine K-SADS symptoms by group. The x-axis is the total score (not including the symptom the graph is describing), this is called the "rest score group"; the y-axis is the mean item response function. If the expectations of IRT are met, then the function should be monotonic. The graphs show that the MDD group had functions that dropped off, usually between the 4th and 5th rest score group; whereas the nonMDD and total sample had monotonic functions. This likely occurred because the item response functions for the total sample and non-MDD group tended to be in the middle range of the depression severity dimension, whereas the MDD group's item response functions were located to the far right of the distribution. These results indicate that parametric IRT models should not be used for the MDD group, but that they are appropriate for the total sample and non-MDD group.

Table A1

Mokken's Loevinger H for Various Groups

	Total Sample		NonMDD Group		MDD Group	
Symptom	*H*	(S.E.)	*H*	(S.E.)	*H*	(S.E.)
Depression/Irritability	0.77	(0.012)	0.29	(0.023)	-0.06*	(0.035)
Pervasive Anhedonia	0.58	(0.011)	0.17	(0.015)	-0.02*	(0.013)
Weight/Appetite Disturbance	0.49	(0.014)	0.15	(0.016)	0.01*	(0.017)
Sleep Disturbance	0.56	(0.012)	0.19	(0.017)	0.02*	(0.011)
Psychomotor Disturbance	0.51	(0.013)	0.21	(0.016)	0.01	(0.015)
Fatigue	0.58	(0.011)	0.21	(0.015)	-0.01*	(0.013)
Self-perceptions	0.53	(0.014)	0.20	(0.019)	0.01*	(0.012)
Cognitive Disturbance	0.60	(0.014)	0.26	(0.018)	0.00*	(0.013)
Suicide	0.49	(0.018)	0.11	(0.020)	0.05*	(0.025)

Note. MDD = major depressive disorder. S.E. = standard error.
**p* < .05.

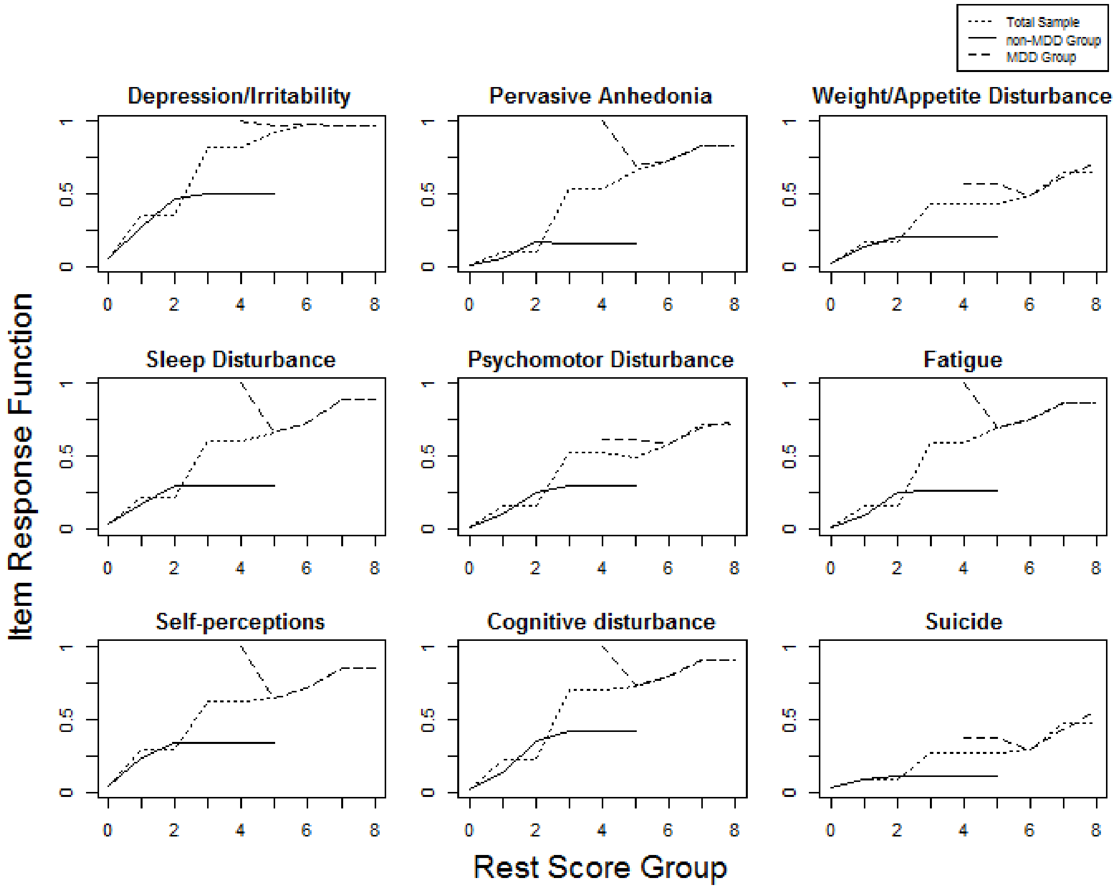

Figure A1. The item response functions (IRF) of the nine K-SADS symptoms by rest score group and sample. The y-axis is the mean IRF and the x-axis is the rest score group.

APPENDIX B

R Code for Analysis in Results Section

Four data sets were used for the analyses in the results section: (a) a data set containing only the nine compound symptoms for the total sample, called “Tot.s9” (b) a data set containing only the nine compound symptoms for the nonMDD group, called “noMDD.s9” (c) a data set containing the seven compound symptoms that did not have DIF and the four component symptoms from the two symptoms that did show DIF for the total sample, called “Tot.s11” and (d) a data set containing the seven compound symptoms that did not have DIF and the four component symptoms from the two symptoms that did show DIF for the nonMDD group, called “noMDD.s11” (see Table B1). In the code below, “DATA_TOTAL” was substituted for either “Tot.s9” or “Tot.s11”, and “DATA__NOMDD” was substituted for “noMDD.s9” and “noMDD.s11” depending on which analysis was being done. Note that the “s9” data sets would be analyzed together; whereas the “s11” data sets would be analyzed together.

Table B1

Data Sets for All Results Section Analyses

	Nine Symptoms		Eleven Symptoms	
Sample / Group Used	Tot.s9	noMDD.s9	Tot.s11	noMDD.s11
Total Sample (*n* = 3,386)	X		X	
NonMDD Group (*n* = 2,246)		X		X
Symptoms Included				
Depression/Irritability	X	X		
Depressed Mood			X	X
Irritability			X	X
Pervasive Anhedonia	X	X	X	X
Weight/Appetite Disturbance	X	X	X	X
Sleep Disturbance	X	X	X	X
Psychomotor Disturbance	X	X		
Psychomotor Agitation			X	X
Psychomotor Retardation			X	X
Fatigue	X	X	X	X
Self-perceptions	X	X	X	X
Cognitive Disturbance	X	X	X	X
Suicide	X	X	X	X

```
R code:
########################################################################
#  Obtaining Item Parameter Estimates for Total Sample & nonMDD Group
#  Equating (difR)
#       Rescaling nonMDD Group’sitem parameters by equal means
#       anchoring
########################################################################
library(difR)
binary.total.2pl.difR <- itemParEst(DATA_TOTAL, model="2PL")
binary.noMDD.2pl.difR <- itemParEst(DATA_NOMDD, model="2PL")
binary.EQUATED.noMDD.2pl.difR <- itemRescale(binary.total.2pl.difR,
                                 binary.noMDD.2pl.difR)
binary.total.2pl.difR
binary.EQUATED.noMDD.2pl.difR

########################################################################
#  Raju’s area DIF method (difR)
#      Performs DIF detection using Raju’s area method
#      raju.nosign is the Z(H)
########################################################################
library(difR)
item.2PL <- rbind(itemParEst(DATA_NOMDD, model="2PL"),
            itemParEst(DATA_TOTAL, model="2PL"))
raju.nosign <- difRaju(irtParam=item.2PL, same.scale=FALSE,
               purify=TRUE, signed=FALSE)
raju.nosign
```

```
#######################################################################
#  Lord’s chi-squared DIF method (difR)
#      Performs DIF detection using Lord’s chi-squared method.
#######################################################################
library(difR)
item.2PL <- rbind(itemParEst(DATA_NOMDD, model="2PL"),
            itemParEst(DATA_TOTAL, model="2PL"))
r.lord <- difLord(irtParam=item.2PL, same.scale=FALSE, purify=TRUE)
r.lord
```

APPENDIX C

R Code for Appendix A Analyses

Three data sets were used for the analyses in Appendix A: (a) a data set with the nine compound symptoms for the total sample, subjects with missing data were deleted because Mokken package cannot analyze dataframes with missing data, this data set was called “Tot.s9.noNA”, (b) a data set with the nine compound symptoms for the nonMDD group, subjects with missing data were deleted, this data set was called “noMDD.s9.noNA”, (c) a data set with the nine compound symptoms for the MDD group, subjects with missing data were deleted, this data set was called “MDD.s9.noNA” (see Table C1).

Table C1

Data Sets for Appendix A Analyses

Sample / Group Used	Tot.s9.noNA	noMDD.s9.noNA	MDD.s9.noNA
Total Sample, no missing data (*n* = 2,590)	X		
NonMDD Group, no missing data (*n* = 1,453)		X	
MDD Group, no missing data (*n* = 1,137)			X
Symptoms Included			
Depression/Irritability	X	X	X
Depressed Mood			
Irritability			
Pervasive Anhedonia	X	X	X
Weight/Appetite Disturbance	X	X	X
Sleep Disturbance	X	X	X
Psychomotor Disturbance	X	X	X
Psychomotor Agitation			
Psychomotor Retardation			
Fatigue	X	X	X
Self-perceptions	X	X	X
Cognitive Disturbance	X	X	X
Suicide	X	X	X

```
R code:
#######################################################################
#  coefH: Scalability coefficents H
#     Computes item-pair scalability coefficents Hij, item
#     scalability coefficents Hi, and scale scalability
#     coefficent H, as well as their standard errors
#  check.monotonicity: Check of Monotonicity
#     Returns a list (of class monotonicity.class) with results
#     from the investigation of monotonicity
#######################################################################

#######################################################################
#  Total Sample (without missing data)
library(mokken)
coefH(Tot.s9.noNA, se = TRUE, nice.output = TRUE, group.var = NULL)

MC_all.noNA <- check.monotonicity(Tot.s9.noNA, minvi = .03, minsize =
               2590/10)
# uses minsize = 2149/10: 2149 represents the non missing subjects from
  the nonMDD group

summary(MC_Tot.noNA)

#######################################################################
#  nonMDD Sample (without missing data)
library(mokken)
coefH(noMDD.s9.noNA, se = TRUE, nice.output = TRUE, group.var = NULL)

MC_noMDD.noNA <- check.monotonicity(noMDD.s9.noNA, minvi = .03, minsize
                 = 1453/10)
# uses minsize = 1452/10: 1452  represents the non missing subjects
  from the nonMDD group

summary(MC_noMDD.noNA)

#######################################################################
#  MDD Sample (without missing data)
library(mokken)
coefH(MDD.s9.noNA, se = TRUE, nice.output = TRUE, group.var = NULL)

MC_MDD.noNA <- check.monotonicity(MDD.s9.noNA, minvi = .03, minsize =
               1137/10)

# uses minsize = 1137/10: 1137 represents the non missing subjects from
  the MDD group

summary(MC_MDD.noNA)
```

REFERENCES

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *The Journal of Abnormal Psychology*, *87*(1), 49–74. doi:http://dx.doi.org/10.1037/0021-843X.87.1.49

Ambrosini, P., & Dixon, J. (1996). *Schedule for Affective Disorders and Schizophrenia for School-Aged Children-Present Version, Version IV-Revised (K-SADS-IV-R)*. Unpublished instrument, Eastern Pennsylvania Psychiatric Institute, Medical College of Pennsylvania, Philadelphia, PA.

Ambrosini, P. J. (2000). Historical Development and Present Status of the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS). *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(1), 49–58. doi:10.1097/00004583-200001000-00016

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

Benazzi, F., Koukopoulos, A., & Akiskal, H. S. (2004). Toward a validation of a new definition of agitated depression as a bipolar mixed state (mixed depression). *European Psychiatry*, *19*(2), 85–90. doi:10.1016/j.eurpsy.2003.09.008

Benazzi, Franco. (2004a). Agitated depression: a valid depression subtype? *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *28*(8), 1279–1285. doi:10.1016/j.pnpbp.2004.06.018

Benazzi, Franco. (2004b). Intra-episode hypomanic symptoms during major depression and their correlates. *Psychiatry and Clinical Neurosciences*, *58*(3), 289–294. doi:10.1111/j.1440-1819.2004.01234.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Cohen, A., & Kim, S. (1993). A Comparison of Lord Chi(2) and Raju Area Measures in Detection of Dif. *Applied Psychological Measurement*, *17*(1), 39–52. doi:10.1177/014662169301700109

Cole, D. A., Cai, L., Martin, N. C., Findling, R. L., Youngstrom, E. A., Garber, J., … Forehand, R. (2011). Structure and measurement of depression in youths: Applying item response theory to clinical data. *Psychological Assessment*, *23*(4), 819–833. doi:http://dx.doi.org/10.1037/a0023518

Dodge, K. A., & Somberg, D. R. (1987). Hostile Attributional Biases among Aggressive Boys Are Exacerbated under Conditions of Threats to the Self. *Child Development*, *58*(1), 213–224. doi:10.2307/1130303

Embretson, S. E., & Hershberger, S. L. (1999). *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., Delbello, M. P., & Soutullo, C. (2001). Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) Mania and Rapid Cycling Sections. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(4), 450–455. doi:10.1097/00004583-200104000-00014

Gladstone, T. R. G., & Kaslow, N. J. (1995). Depression and attributions in children and adolescents: A meta-analytic review. *Journal of Abnormal Child Psychology*, *23*(5), 597–606. doi:10.1007/BF01447664

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., … Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(7), 980–988. doi:10.1097/00004583-199707000-00021

Kaufman, J., Birmaher, B., Brent, D., RAO, U., & Ryan, N. (1996). *Kiddie SADS- Present and Lifetime Version (K-SADS-PL)*. Unpublished instrument, Western Psychiatric Institute and Clinics, University of Pittsburg School of Medicine, PA.

Leventhal, A. M., Pettit, J. W., & Lewinsohn, P. M. (2008). Characterizing major depression phenotypes by presence and type of psychomotor disturbance in adolescents and young adults. *Depression and Anxiety*, *25*(7), 575–592. doi:10.1002/da.20328

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets and Zeitlinger.

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

Orvaschel, H. (1994). *Schedule for Affective Disorders and Schizophrenia for School-Age Children-Epidemiological Version* (5th ed.). Ft. Lauderdale, FL: Nova Southeastern University.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. doi:10.1007/BF02294403

Raju, N. S. (1990). Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions. *Applied Psychological Measurement*, *14*(2), 197–207. doi:10.1177/014662169001400208

Schuur, W. H. van. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Political Analysis*, *11*(2), 139–163. doi:10.1093/pan/mpg002

Steinberg, M. S., & Dodge, K. A. (1983). Attributional Bias in Aggressive Adolescent Boys and Girls. *Journal of Social and Clinical Psychology*, *1*(4), 312–321. doi:http://dx.doi.org/10.1521/jscp.1983.1.4.312

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.