MAKING DIAGNOSTIC THRESHOLDS LESS ARBITRARY

A Thesis

by

ALEXIS ARIANA UNGER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2011

Major Subject: Psychology

Making Diagnostic Thresholds Less Arbitrary

Copyright 2011 Alexis Ariana Unger

MAKING DIAGNOSTIC THRESHOLDS LESS ARBITRARY

A Thesis

by

ALEXIS ARIANA UNGER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Steve Balsis
Committee Members,	Aaron Taylor
	Rick Peterson
Head of Department,	Ludy Benjamin

May 2011

Major Subject: Psychology

ABSTRACT

Making Diagnostic Thresholds Less Arbitrary. (May 2011)

Alexis Ariana Unger, B.S.; B.S., Michigan State University

Chair of Advisory Committee: Dr. Steve Balsis

The application of diagnostic thresholds plays an important role in the classification of mental disorders. Despite their importance, many diagnostic thresholds are set arbitrarily, without much empirical support. This paper seeks to introduce and analyze a new empirically based way of setting diagnostic thresholds for a category of mental disorders that has historically had arbitrary thresholds, the personality disorders (PDs). I analyzed data from over 2,000 participants that were part of the Methods to Improve Diagnostic Assessment and Services (MIDAS) database. Results revealed that functional outcome scores, as measured by Global Assessment of Functioning (GAF) scores, could be used to identify diagnostic thresholds and that the optimal thresholds varied somewhat by personality disorder (PD) along the spectrum of latent severity. Using the Item response theory (IRT)-based approach, the optimal threshold along the spectrum of latent severity for the different PDs ranged from $\theta = 1.50$ to 2.25. Effect sizes using the IRT-based approach ranged from .34 to 1.55. These findings suggest that linking diagnostic thresholds to functional outcomes and thereby making them less arbitrary is an achievable goal. This study has introduced a new and uncomplicated way to empirically set diagnostic thresholds while also taking into consideration that items

within diagnostic sets may function differently. Although purely an initial demonstration meant only to serve as an example, by using this approach, there exists the potential that diagnostic thresholds for all disorders could one day be set on an empirical basis.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Steve Balsis, and my committee members, Dr. Aaron Taylor, and Dr. Rick Peterson, for their guidance and support throughout the course of this research.

In addition, I would like to lend my thanks to my friends and colleagues, as well as the department faculty and staff, for making my time at Texas A&M University such a rewarding and memorable experience. I would also like to extend my gratitude to Dr. Mark Zimmerman from the Department of Psychiatry and Human Behavior at the Brown University School of Medicine for allowing me access to his data, and for his collaboration and feedback.

Finally, I would like to thank to my family and friends for their encouragement, unwavering support, and love throughout this process. I owe them more than words can express.

NOMENCLATURE

2PL	Two-Parameter Logistic
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision
GAF	Global Assessment of Functioning
ICC	Item Characteristic Curve
IRT	Item Response Theory
MAP	*maximum a posteriori*
MIDAS	Methods to Improve Diagnostic and Assessment Services Database
PD/PDs	Personality Disorder/Personality Disorders
SIDP-IV	Structured Interview for DSM-IV Personality

TABLE OF CONTENTS

1. INTRODUCTION

In the Diagnostic and Statistical Manual for Mental Disorders, 4th edition, text revision (DSM-IV-TR), a diagnostic threshold for a mental disorder is a cut-point or cutoff that indicates the minimum number of diagnostic criteria necessary to be given a diagnosis. Thresholds are used to determine whether or not an individual meets enough of the criteria to be eligible for a diagnosis. If a person does not meet enough criteria to place them at or above a given threshold, a diagnosis cannot be given. Thus it follows that thresholds are fundamental to the identification of mental disorders.

The application of thresholds can be seen in a variety of situations, including clinical practice, assessment and treatment research, communication among professionals, law, and public policy. In everyday clinical practice diagnostic thresholds are used to make treatment decisions. For instance, the treatment selected for a child typically differs depending on whether he or she receives a diagnosis of conduct disorder. Moreover, if a client no longer meets criteria for a given diagnosis during treatment (going from above threshold to below threshold) a clinician may decide to terminate treatment. In assessment research, thresholds are used to determine the reliability and validity of a given measure for the mental disorder or construct that it measures. For example, a questionnaire may be studied to determine its interrater reliability, applying diagnostic thresholds to help make decisions about agreement among raters. Further, in treatment outcome research, thresholds are used to help

---

This thesis follows the style of *Journal of Personality Disorders.*

demonstrate the efficacy and effectiveness of the intervention being examined. As an example, in some research designs, when researchers attempt to measure whether or not an individual has "recovered", or moved from "disordered" to "non-disordered" over the course of a given treatment, thresholds are regularly applied. If a significant number of individuals who previously were above threshold are now below threshold post-treatment, the treatment can be considered efficacious. In addition, diagnostic thresholds play an important role in facilitating communication among professionals by providing a common language with which to describe individuals. Certain characteristics are associated with individuals who are "above threshold" and thus meet criteria for any given diagnosis, versus those who are not. As an example, someone who is diagnosed with mental retardation and meets criteria will typically have an IQ of approximately 70 or below and display deficits or impairments in functioning (American Psychiatric Association, 2000). Within this diagnostic category, there are gradations of mild, moderate, severe, and profound mental retardation, each with a threshold created by a range of IQ scores. It follows that each threshold or gradation of mental retardation is associated with varying characteristics. It is also common for legal decisions to make use of thresholds (Buchanan, 2005; Edens & Petrila, 2006). For example, in the state of Washington, sexual violent predator laws require that an individual committing a sexual offense can only be held in custody if he or she suffers from a 'mental abnormality or personality disorder' (Buchanan, 2005). Finally, thresholds heavily influence public policy. As an example, to receive reimbursement from health insurance and third party payers for services and treatments rendered, an individual must have received a

diagnosis. The diagnosis is based, at least in part, on the person being above a given threshold for a disorder. In this domain, thresholds also influence an individual's ability to qualify for access to programs such as special education. Thus, given their importance and widespread use described here, there are many practical reasons to analyze diagnostic thresholds.

Despite the fact that diagnostic thresholds are very important and widely used, there are multiple problems with the current practice of setting and applying them. One problem with diagnostic thresholds is that they have oftentimes been set in a rather arbitrary manner. As an example of this arbitrariness, consider the personality disorders (PDs). In the DSM-IV-TR, there are 10 PDs, each of which has a specified number of criteria, which varies from seven to nine (American Psychiatric Association, 2000). To be diagnosed with one of these PDs, an individual must be at or above the diagnostic threshold, which means meeting a minimum specified number of criteria. This diagnostic threshold varies across the spectrum of PDs. For example, to meet criteria for antisocial PD, an individual must endorse three out of seven criteria; for narcissistic PD, a person must meet five out of nine criteria (American Psychiatric Association, 2000). Although these thresholds vary across disorder, many of them are not empirically based. In fact, seven of them were chosen on an almost entirely arbitrary basis, with only borderline, schizotypal, and antisocial PD thresholds being based even in part on empirical evidence (Widiger & Trull, 2007; Spitzer, Endicot, & Gibbon, 1979; Widiger et al., 1996). Research backing the chosen diagnostic thresholds for the remaining PDs is essentially nonexistent. Thus it becomes imperative to find empirical evidence to

support the remaining chosen diagnostic thresholds. Although this task may at first seem rather simple, a further issue with the current thresholds makes this process more complicated.

A second issue with the current diagnostic thresholds is that they are based on raw scores. In other words, an individual simply needs to meet a certain number of criteria to be diagnosed with a given disorder, and which items are endorsed is not taken into account. Each criterion counts as one "point" toward the final raw score for diagnosis. In previous research, this method of using raw scores and therefore weighting each criterion equally has been shown to yield an imprecise measure of the latent disorder construct being studied (Cooper & Balsis, 2009). This imprecision occurs because although each criterion is weighted equally, each actually contributes differentially to the underlying level or latent level of pathology, with some criteria being more closely related to the underlying construct being measured and some criteria measuring different levels of severity. For example, consider the diagnostic criteria for antisocial PD. One criterion is "impulsivity or failure to plan ahead" and another is "irritability and aggressiveness, as indicated by repeated physical fights or assaults" (American Psychiatric Association, 2000). Each of these items tells something slightly different about the underlying latent trait of antisocial PD. The item assessing physical fights and assaults can be thought of as a more "difficult" or "severe" item than the item assessing impulsivity, because a person with more antisocial tendencies has a much higher probability of endorsing the physical fights and assaults item than a person with lower levels of these tendencies. In the same sense, the item assessing impulsivity can be

thought of as an "easier" item as it is likely that even a person with very little underlying antisocial tendency might endorse this item as well as a person who has high levels of antisocial PD.

This example nicely demonstrates that each criterion for a given disorder differs in the severity of the underlying pathology, in this case antisocial PD, which it is measuring. Moreover, as noted above, each criterion is also differentially related to the underlying construct being measured. In this example, the item about physical fights and assaults is likely to be more strongly linked to antisocial behavior. In other words, it is highly likely that a person with this symptom is experiencing it due to an antisocial tendency and not some other underlying trait. By contrast, someone with impulsivity, the "easier" item, could be experiencing this symptom not only due to antisocial PD, but possibly as a product of borderline PD, attention-deficit/hyperactivity disorder (ADHD), or as part of a substance abuse problem. Thus the "easy" item may be less related to the latent construct of antisocial PD. Under the current diagnostic system, which uses raw scores to determine diagnoses, each of these two items is given equal weight, despite the fact that they measure different levels of antisocial PD severity and are differentially related to the construct.

Ideally, then, thresholds should not only be empirically-based, they should also take into consideration the fact that each item may measure the underlying construct differently. Item response theory (IRT) grants the statistical ability both to consider how items function and to empirically determine diagnostic thresholds. IRT considers how items function by estimating parameters for each item. In a commonly used IRT model,

the two parameter logistic model (2PL), there are two such parameters, the *a* (discrimination) and *b* (threshold) parameters. In this case, the discrimination parameter indicates how strongly a given criterion (item) is related to the underlying level of the PD being measured. The threshold parameter indicates where on the underlying construct of the PD being measured an item best discriminates. Taken together, these parameters tell both the difficulty and how related a criterion is to the latent variable, theta (θ). IRT also allows one to determine where individuals fall on this latent dimension, based on their responses to a set of diagnostic criteria.

Historically, IRT has been employed primarily in educational testing to estimate academic abilities (Hambleton, Swaminathan, & Rogers, 1991; Baker, 2001). More recently, it has also been used to estimate verbal and mathematical ability on the SAT and GRE. IRT also lends itself to the estimation of underlying, pathology using sets of diagnostic criteria and is well-suited for determining how each diagnostic criterion in the DSM for a given mental health construct is related to the underlying level of pathology for that construct.

Even given the use of IRT scoring to more precisely estimate each individual's level of pathology the question of where or how should we draw the threshold remains open. In this paper, I demonstrate a new approach to choosing thresholds for making diagnoses. For illustrative purposes, I demonstrate application of this method only for the PDs. The DSM-IV-TR PDs are a useful category of mental disorders in which to begin to examine new ways of setting and applying diagnostic thresholds for several reasons. One, all PDs are fully polythetic, a characteristic that makes examining the

influence of individual items simpler. Two, as mentioned above, many of the PDs have thresholds that historically have not been empirically based. Three, the DSM PD items (diagnostic criteria) have already been examined within a diagnostic threshold framework (Balsis, Lowmaster, Cooper, & Benge, in press; Cooper & Balsis, 2009; Cooper, Balsis, & Zimmerman, 2010). More simply stated, it has already been shown that setting diagnostic thresholds is a problem for these disorders. For this category of mental disorder, it follows that one possible option for drawing diagnostic thresholds would be to draw them at the same level across all of the PDs, say at 2 *SD*s above the mean of latent pathology. This method of selecting a diagnostic threshold, however, would still be lacking empirical support.

Ideally, thresholds would not be arbitrary but would instead be chosen for each disorder on some sort of empirical basis. There are several possible empirical approaches one could take; however, a self-evident approach would be to base the threshold on the ability of diagnoses made using it to predict important functional outcomes. One such functional outcome is the Global Assessment of Functioning (GAF) scale (American Psychiatric Association, 2000), which provides a comprehensive method of assessing an individual's overall level of functioning and is already a part of the DSM system. Using GAF as an outcome measure for identifying thresholds might be both logical and useful.

The goal of this thesis is thus to determine, for each PD, where along the continuum of latent pathology is the best place to draw the diagnostic threshold. The criterion for choosing a best threshold is predictive validity of the diagnosis for

functional impairment as measured by the GAF. In this way, I hope to be able to provide a suggested diagnostic threshold for each PD that will be based on a more empirical method than has been previously done. It is important to note that my goal for this thesis is not to try to create definitive cutoffs that should be used in place of the current DSM thresholds, but rather to offer a first look at a more systematic method of determining diagnostic thresholds.

2. METHOD

2.1 *Participants*

The data for these analyses comes from the Methods to Improve Diagnostic Assessment and Services (MIDAS) database, which includes over 2,000 participants. They were outpatients at a university clinic over the past 12 years. These individuals were on average 38.5 years old (*SD*= 13.0) and most had earned at least a GED or high school diploma (91%). The majority were female (61%, *n* = 1,818) and White (87%, *n* = 2,622). Other ethnic identifications included Black (5%, *n* = 135), Hispanic (3%, *n* = 77), Asian (1%, *n* = 28), Portuguese (<1%, *n* = 96), American Indian (<1%, *n* = 1), and Other (1%, *n* = 41). Informed consent was obtained from all participants.

Analyses were conducted separately for each PD. Each analysis included a slightly different number of participants because individuals were excluded from the analysis for a particular PD if they had failed to answer any of the items on the *Structured Interview for DSM-IV Personality* (SIDP-IV, Pfohl, Blum, & Zimmerman, 1997) for that PD. Participants were excluded from all analysis if they did not have a GAF score recorded. The total number of participants included in the analyses for each PD as follows: 2,673 for antisocial PD, 2,149 for avoidant PD, 2,869 for borderline PD, 2,145 for dependent PD, 2,145 for histrionic PD, 2,143 for narcissistic PD, 2,142 for obsessive-compulsive PD, for 2,150 paranoid PD, 2,148 for schizoid PD, and 2,135 for schizotypal PD. The gender and ethnic breakdown of the sample did not vary significantly across the different analyses.

2.2 *Measures*

The SIDP-IV was used for this study. This instrument is a semi-structured interview that assesses for 13 PDs, but this study only analyzed the 10 PDs that are included in the DSM-IV (American Psychiatric Association, 1994). Each item on the SIDP-IV is created from a corresponding DSM-IV diagnostic criterion for a given PD. The responses for each item range from 0 (meaning the criterion is "not present") to 3 (meaning the criterion is "strongly present"). This measure has been shown to demonstrate good interrater reliability (Jane, Pagan, Turkheimer, Fiedler, & Oltmanns, 2007), even when it is administered by individuals who do not have a strong background in assessment (Balsis, Eaton, Zona, & Oltmanns, 2006). In this sample, SIDP-IV administration was conducted by individuals who had received thorough training and who were subject to supervision throughout the process of data collection.

2.3 *Data Analyses*

Analyses were performed using IRT to first estimate participants' scores along the underlying dimension for each of the 10 PDs, and then to evaluate a series of candidate cutpoints for each PD to find one that maximally separated participants on their GAF scores. These optimal cutpoints based on IRT and estimation of latent values of θ for each participant were then compared to the existing cutpoints used with SIDP-IV raw scores to discover whether the new cutpoints would outperform the existing cutpoints in defining groups that differed most on GAF.

The first step in my analyses was to dichotomize responses. The purpose of dichotomizing these responses was to allow each response to be interpreted in terms of

the DSM-IV, where an individual either endorses or does not endorse any given item. If participants originally gave a response of 0 ("not present") or 1 ("subthreshold"), their responses were recoded into 0s, and if they gave a response of 2 ("present") or 3 ("strongly present") their responses were recoded into 1s. Thus for a given PD, the total number of 1s that an individual participant had was equivalent to the number of diagnostic criteria met based on the DSM-IV.

The next step was to use IRT to estimate each participant's score along the latent dimension (θ) for each PD by considering both how the items function and the scores obtained on the items. These analyses were run using MULTILOG software (Thissen, 1991). The 2 parameter logistic (2PL) model was used to estimate the probability of obtaining a given score for each PD at each level of θ. The *a* (discrimination) and *b* (threshold) parameters were then estimated for each of the criteria for each PD (please see the introduction for a more thorough description of the *a* and *b* parameters). Finally, each participant's θ score for each PD was estimated by taking into consideration not only how the diagnostic criteria for a given PD function (i.e., their *a* and *b* parameters) but also the participant's response to each of the criteria (present versus not present). The θ scores were estimated using *maximum a posterioris* (MAPs, Thissen & Orlando, 2001). MAPs are simply posterior estimates of θ that take into consideration what was just described above: a combination of how items function and a person's response to each item.

Due to the fact that MAPs are so central to my statistical analyses I here provide an example to demonstrate how they are estimated. The exposition here closely follows

that of Balsis, Unger, Benge, Geraci, and Doody (2010; see also Thissen & Orlando, 2001). I consider again the example of two criteria for antisocial personality disorder; the impulsivity item and the physical fights and assaults item. The "easier" item, the impulsivity item, might have a *b*, or difficulty parameter, of -1.00, whereas the more "difficult" item, the physical fights and assaults item, might have a difficulty parameter of 2.00. Because an item's difficulty (or location) parameter tells the value of θ at which a person has a 50% probability of endorsing the item, an individual 1 *SD* below the mean on latent antisocial PD severity (θ = -1.00) has a 50% probability of endorsing the criterion for impulsivity, but a much lower probability of endorsing the more "difficult" physical fights and assaults criterion. By contrast, an individual 2 standard deviations above the mean on severity of antisocial PD (θ = 2.00) will endorse the physical fights and assaults criterion with 50% probability, and a much higher probability of endorsing the "eaiser" impulsivity item. Thus, the difficulty parameter is showing that these two criteria differ in their "difficulty", with the impulsivity item being less "difficult", because more people, even with less severe disorders, will endorse it.

As described in the introduction, these two criteria not only differ in their degree of "difficulty", but also with the degree to which they are associated with the construct of antisocial PD. The impulsivity criterion is less closely related to the construct of antisocial PD than is the physical fights and assaults criterion. This difference in degree of relatedness is captured by the *a* parameters, which are simply slopes of the item function curves of the difficulty parameter at a level of 50% probability of endorsement.

Steeper slope values are indicative of items that are more strongly related to the underlying construct being measured, in this case, antisocial PD.

In graphing these two parameters, Item Characteristic Curves (ICCs) can be generated that show the probability of any given item being endorsed across the entire spectrum of the latent trait. In this case, the ICCs would signify the probability that the physical fights and assaults item and the impulsivity item will be endorsed across the entire spectrum of antisocial PD. Given the information provided above, in this example, you can imagine that the physical fights and assaults item, the item that is both "harder" and more closely linked to antisocial PD would have a curve that is not only shifted to the right of the curve generated by the impulsivity item, but it would also have a steeper slope.

I have thus shown how any particular criterion or item that is endorsed can be modeled for a participant. It is also possible to model the probability that an item is *not* endorsed. This probability is simply 1 minus the probability that the criterion is endorsed. As an example, if the impulsivity item has a 80% chance of being endorsed at $\theta = .25$ *SD*s of antisocial PD pathology, then there is also a 20% probability that a person with $\theta = .25$ *SD*s will not endorse the item. The curve generated by these values is termed the inverse ICC. So for this simple example, two curves could be plotted: one showing the probability of endorsing the impulsivity item but not endorsing the physical fights and assaults item and another curve displaying non-endorsement of the impulsivity item but a positive endorsement of the physical fights and assaults item. In addition, the Gaussian distribution can be plotted as a basis of comparison. In

generating a MAP score, in this case estimating the amount of latent antisocial PD severity, it is these three curves that are considered.

In order to calculate the joint probability that a given response pattern will occur, for example endorsement of the impulsivity criterion and non-endorsement of the physical fights and assaults criterion, the ICC for the item endorsed and the inverse ICC for the item not endorsed are multiplied together at each value along the spectrum of θ, or underlying antisocial PD pathology. The next step assumes a normal distribution of θ and is to create a maximum likelihood estimation function for estimating the MAP by multiplying this joint probability by the normal distribution.

Oftentimes, the mode of this line, the MAP, is used as an estimate of θ, in this case antisocial PD. Unfortunately, my estimation of underlying personality pathology with the SIDP-IV is much more complicated because, instead of considering how 2 items function as in the example, estimation of participants' underlying PD pathology, must consider how all of the items, or diagnostic criteria, for that PD function. So for example, for borderline PD, there are a total of 9 criteria to consider. The complexity of this study is further increased because the calculations described must be performed for all 10 of the PDs being examined. MULTILOG software is able to perform all these calculations when estimating MAPs.

Once MAPs were estimated for each participant on each PD, a number of candidate cutpoints were drawn along the range of θ for each PD, and were compared in terms of their ability to maximally separate participants on their GAF scores. In a separate analysis for each PD, participants were sorted by their level of θ, ranging from

smallest to largest. The values of θ across the 10 PDs for this sample ranged from -.16 to 3.10. As an example, for schziotypal PD, the values for θ ranged from -.16 to 2.76. The next step in this analysis was to draw a number of candidate cutpoints along the spectrum of θ for each PD. So once again using schizotypal PD as an example, candidate cutpoints were drawn along the available spectrum of θ from -.16 to 2.76 in increments of .25 *SD*. I chose increments of this size because it seemed to be a width small enough to capture differences in effect size without being so large that it would miss better alternative cutpoints in between the candidate cutpoints. Because the cutpoints would be used to estimate effect sizes, it was necessary that there be sufficient participants both above and below each candidate cutpoint to allow for the differences in effect size to be calculated. So for instance, when dealing with schizotypal PD, the first cutpoint was drawn at a θ value of 0 rather than -.25 because there were no participants who had a θ value below -.16.

From here, the next step was to calculate an effect size for each candidate cutpoint so the optimal cutpoint with the largest effect size could be found. These calculations were repeated for each candidate cutpoint for each of the 10 PDs. The effect sizes were calculated using the participants' GAF scores. The formula was that of an independent samples t-test:

$$t = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{s_B^2(n_B - 1) + s_A^2(n_A - 1)}{n_B + n_A - 2}}}$$

B and A in the formula refer to groups below and above the candidate cutpoint. Once again using the schizotypal PD example, for the effect size calculation at θ = 2.25 *SD*, the *n* above the threshold was 12 because there were 12 participants with a θ greater than or equal to 2.25 and the *n* below the threshold was 2,123, because there were that many participants who had a θ less than 2.25. The average GAF score above the candidate cutpoint was 39.3 (*SD*= 9.55) and the average score below was 53.9 (*SD*= 9.42). Thus the effect size for this candidate cutpoint was calculated to be 1.53.

The following step was to determine the best, or most predictive, threshold for each PD. This decision was reached by choosing the threshold where the effect size was the largest. However, effect sizes were not considered if the number of participants on either side of the candidate cutpoint used to generate the effect size was less than 10. Effect sizes can become easily inflated with small values of *n*, and thus if drawing a certain threshold would result in there being only 2 cases above this threshold, I would move backwards along the spectrum of θ to the next closest .25 increment. As an example, for schizotypal PD, the cutpoint was drawn at a θ value of 2.25. There were 12 participants with θ values at or above 2.25, including values of 2.33, 2.41, 2.42, 2.46, 2.48, 2.52, 2.55, 2.55, 2.62, 2.63, 2.67, 2.73, and 2.76. At first glance, it might seem like, given the original rule of drawing cutpoints at θ increments of .25 that additional cutpoints should be drawn at 2.50 and 2.75. However, if a cutpoint were to be drawn at 2.50, this would include only 8 participants, and one drawn at 2.75 would contain only one case above the threshold. Thus it follows, that neither of these cutpoints would contain at least 10 participants above the threshold, so for the purpose of these analyses,

the final cutpoint for schizotypal PD was drawn at a θ value of 2.25. As a result of this process, for this example, the most effective threshold for schizotypal PD was 2.25. This process was repeated for each PD.

Finally, the optimal thresholds chosen using this method were compared to the existing DSM-IV-TR cutpoints. For each PD, the participants' raw total scores on the SIDP-IV were calculated. This was done by summing the recoded items described above (scored 0 and 1). So for example, if, after recoding, a participant's responses for antisocial PD were 0 0 0 0 1 1 1, their total score would be 3. From here, a cutpoint was drawn for each PD at the level where the current DSM-IV-TR threshold lies. So once again for antisocial PD, the threshold was drawn at a total score of 3, because a person must endorse 3 out of the 7 criteria in order to be diagnosed with antisocial PD. Using these thresholds, effect sizes for the GAF were calculated as they were for the candidate cutpoints drawn at different levels of θ. Finally, a comparison was made for each PD between the effect size based on the DSM-IV-TR threshold and the effect size based on the optimal threshold on the θ scale, found as described above.

3. RESULTS

Results indicated that GAF scores could be used to determine diagnostic thresholds and that the optimal thresholds varied somewhat by disorder (see Table 1). Using the IRT-based approach, the optimal threshold along the spectrum of latent severity for the different PDs ranged from $\theta = 1.50$ to 2.25, with antisocial, narcissistic, and obsessive-compulsive PDs having an optimal threshold at $\theta = 1.50$, borderline, histrionic, and avoidant PDs having an optimal threshold at $\theta = 2.00$, and all of the cluster A PDs and dependent PD having an optimal threshold drawn at $\theta = 2.25$. Effect sizes using the IRT-based approach ranged from .34 for obsessive-compulsive PD to 1.55 for schizotypal PD.

For purposes of comparison, I also calculated effect sizes using the current DSM approach. The DSM-based effect sizes were calculated by drawing the threshold at the minimum number of items necessary listed in the DSM-IV-TR to meet criteria. For 7 of the 10 PDs (paranoid, schizoid, histrionic, narcissistic, and the cluster C PDs), the effect size generated using the IRT approach was greater than the effect size for the classic DSM approach (see Table 1). For one other PD, schizotypal PD, the effect size was the same for both methods, and for two PDs, borderline PD and antisocial PD, the effect size calculated using the DSM approach was similar to, but slightly higher than, the effect size for the IRT approach. The largest difference in effect size between the two methods was found for paranoid PD, with a sizeable gap of .41 in favor of the IRT approach.

TABLE 1. Diagnostic thresholds across method and associated effect sizes

	IRT-approach		DSM approach	
	Threshold	Effect size	Threshold	Effect size
Cluster A				
Paranoid	2.25	1.29	4	0.88
Schizoid	2.25	1.20	4	0.91
Schizotypal	2.25	1.55	5	1.55
Cluster B				
Antisocial	1.50	0.67	3	0.69
Borderline	2.00	1.01	5	0.97
Histrionic	2.00	0.61	4	0.72
Narcissistic	1.50	0.45	5	0.39
Cluster C				
Avoidant	2.00	0.83	4	0.66
Dependent	2.25	1.09	5	0.77
Obs-Comp	1.50	0.34	4	0.32

4. DISCUSSION AND CONCLUSIONS

Thresholds continue to play a crucial role in the classification of mental disorders and have many important clinical applications. However, as discussed here, current diagnostic thresholds are problematic because they are not only chosen in an arbitrary manner, but also because they are based on raw scores. I have shown in this paper that by using IRT scoring in conjunction with a functional outcome measure, I can both account for the fact that diagnostic criteria function differently and set thresholds empirically. The findings of this study have potentially far-reaching and important applications. Although I have used PDs as a simple example of how to empirically set diagnostic thresholds, the methodology used here to create empirically based thresholds could easily be applied to a host of other disorders and functional outcomes.

Although this study does have important possible implications, it is not without limitations. First, the sample studied was largely white, relatively young, and made up entirely of outpatients. Results may therefore not generalize well to other demographic groups. Additionally, although the sample was large, it did not include participants who covered the full range of possible values of θ for each disorder. The GAF is also less than ideal as a functional outcome measure because of its sometimes-questioned reliability (Goldman, Skodol, & Lave, 1992; Rey, Plapp, Stewart, Richards, & Bashir, 1987). At the same time, it is important to consider that my purpose was simply to use this measure as an exemplar. Future research could apply the same approach described here with any chosen functional outcome of interest.

Finally, another challenge associated with using an outpatient sample for this type of analysis is that the magnitude of the effect sizes is likely attenuated by at least two factors (Cooper, Balsis, & Zimmerman, 2010). One, the range of GAF scores observed in outpatient samples is naturally truncated. Although GAF scores can range from 1 to 100, patients with GAF scores in an outpatient sample much below 40 or much above 80 are extremely rare, and the resultant truncation leads to an artificially low correlation between GAF and each latent dimension. Two, in an outpatient sample, GAF scores are often affected by factors other than PD trait levels in such a way that masks the true relationship between GAF and each PD dimension. For example, if an outpatient only has small amounts of one PD trait (say, narcissism), he or she is still likely to have poor functioning (and hence a low GAF score) for another reason, such as the presence of depression. In other words, outpatients who are low or high on any one PD trait are likely to be experiencing significant dysfunction. Because outpatients are experiencing dysfunction regardless of their standing on any particular dimension, the correlation between any particular PD and GAF may be artificially low. To comprehensively and adequately study the relationship between GAF and any mental disorder construct, it would be best to isolate a sample of individuals who span the full GAF spectrum and who are uniquely high or low on a particular PD. It is important to note that, even though this sample is imperfect for exploring issues related to functioning, all correlations between latent PD dimensions and GAF were statistically significant.

In total, this study has introduced a new and uncomplicated way to empirically set diagnostic thresholds while also taking into consideration that items

within diagnostic sets may function differently. Using this approach, diagnostic thresholds for all disorders may be set on an empirical basis. I reiterate that this study is purely an initial demonstration of a method of setting thresholds, meant to serve as an example, rather than an argument that the particular thresholds found in this study should be adopted in favor of existing ones. I also readily concede the limitation of thresholds in general: I would not argue that if a threshold is drawn at 2.00 *SD*s of latent pathology, that a person with 1.99 *SD*s of pathology is radically different from a person with 2.01 *SD*s of pathology. As discussed above, though, given that thresholds are so widely used, they must be drawn somewhere, and the method of choosing optimal thresholds presented in this paper is superior to existing methods because it allows them to be drawn empirically. It is quite possible, that in the future, with a more sophisticated classification system, that we may find that there are actually multiple thresholds for different outcomes. A more sophisticated system with multiple thresholds linked to various outcomes could potentially help our field make substantial gains in the efficiency and effectiveness of our interventions.

REFERENCES

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

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

Baker, F.B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Balsis, S., Eaton, N.R., Zona, D. M., & Oltmanns, T.F. (2006). Teaching advanced psychopathology: A method that promotes basic undergraduate clinical and research experience. *Teaching of Psychology, 33,* 242-245.

Balsis, S., Lowmaster, S., Cooper, L., & Benge, J.F. (In press). Personality disorder diagnosticthresholds correspond to different levels of latent pathology. *Journal of Personality Disorders.*

Balsis,S., Unger, A.A., Benge, J.F., Geraci L., & Doody, R.S. (2010). *Gaining Precision on the ADAS-cog: A Comparison Between Item Response Theory-Based Scores and Raw Scores*. Manuscript submitted for publication.

Buchanan, A. (2005). Descriptive diagnosis, personality disorder and detention. *Journal of Forensic Psychiatry & Psychology, 16*, 538-551.

Cooper, L.D., & Balsis, S. (2009). When less is more: How fewer diagnostic criteria can indicate greater severity. *Psychological Assessment, 21*, 285-293.

Cooper, L.D., Balsis, S., & Zimmerman, M. (2010). Challenges associated with a polythetic diagnostic system: Criteria combinations in personality disorders.

Journal of Abnormal Psychology, 119, 886-895.

Edens, J.F., & Petrila, J. (2006). Legal and ethical issues in the assessment and treatment of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp.573-588). New York: Guilford.

Goldman, H.H., Skodol, A.E., & Lave, T.R. (1992). Revising Axis V for DSM-IV: A review of measures of social functioning. *American Journal of Psychiatry, 149*, 1148-1156.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Jane, J.S., Pagan, J.L., Turkheimer, E., Fielder, E.R., & Oltmanns, T.F. (2006). The interrater reliability of the Structured Interview for DSM-IV Personality. *Comprehensive Psychiatry, 47*, 368-375.

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured interview for DSM-IV personality*. Washington, DC: American Psychiatric Press.

Rey, J.M., Plapp, J.M., Stewart, G., Richards, I., & Bashir, M. (1987). Reliability of the psychosocial axes of DSM-III in an adolescent population. *British Journal of Psychiatry, 150*, 228-234.

Spitzer, R.L., Endicott, J., & Gibbon, M. (1979). Crossing the border in borderline personality and borderline schizophrenia. *Archives of General Psychiatry, 36*, 17-24.

Thissen, D. (1991). *MULTILOG user's guide* (Version 6). Chicago: Scientific Software.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two

categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 73-140). Hillsdale, NJ: Erlbaum.

Widiger, T.A., Cadoret, R., Hare, R., Robins, L., Rutherford, M., Zanarini, M., et al. (1996). *DSM-IV* antisocial personality disorder field trial. *Journal of Abnormal Pychology*, *105*, 3-16.

Widiger, T.A., & Trull, T.J. (2007). Plate tectonics in the classification of personality disorder:shifting to a dimensional model. *American Psychologist*, *62*, 71-83.

VITA

Alexis Ariana Unger received her Bachelor of Science degree in psychology from Michigan State University in 2009. She also received a Bachelor of Science degree in human biology from Michigan State University in 2009, as well as a specialization in bioethics, humanities, and society. She entered the Clinical Psychology program at Texas A&M University in August 2009 and received her Master of Science degree in May 2011. Her main research interest is in the area of assessment within clinical psychology. Much of her research focuses on the assessment of disorders using Item Response Theory. One line of research is aimed at improving the measurement and diagnosis of personality disorders. The goal of this research is to improve the precision in the measurement of psychopathology. A second line of research focuses on functional outcomes in individuals with personality pathology. Ms. Unger may be reached at Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235. Her email is ungerale@tamu.edu.