2014

Diagnostic Accuracy Of Self-Report Instruments In A Non-Clinical Sample: A Receiver And Operating Characteristics (Roc) Analysis

Lindsay Trent
University of Mississippi

DIAGNOSTIC ACCURACY OF SELF-REPORT INSTRUMENTS IN A NON-CLINICAL SAMPLE: A RECEIVER AND OPERATING CHARACTERISTICS (ROC) ANALYSIS

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
The University of Mississippi

By

LINDSAY R. TRENT, M.A.

JULY 2014

ABSTRACT

INTRODUCTION: Informed by previous research on structured interviewing, diagnostic reliability, signal detection theory, and the need to enhance accurate assessment in clinical practice, this study examines the diagnostic utility of a wide cross-section of self-report instruments assessing symptoms of psychopathology in terms of predicting diagnostic classification. METHODS: The reference criterion (i.e., "actual" diagnostic classification) was derived from responses to research-based semi-structured psychiatric interviews (i.e., Anxiety Disorders Interview Schedule for DSM-IV and the Eating Disorders module of the Structured Clinical Interview for the DSM-IV). Nineteen self-report measures were subjected to a Receiver Operating Characteristics (ROC) analysis using a range of diagnoses and diagnostic statuses as the reference criterion for prediction. Standard ROC indices of predictive validity (i.e., Area Under ROC Curve, sensitivity, specificity, efficiency, etc.) are reported for diagnostic referents with equal importance placed on sensitivity and specificity (i.e., *kappa* = .5). Non-parametric methods were utilized for comparison of measure performance within diagnostic categories accounting for sample size and bivariate correlation between measures. RESULTS: As expected, significant correlations routinely emerged between measures and ROC analyses revealed significant overlap in terms of measures' predictive validity across diagnostic categories (i.e., given overlap of symptoms and psychological constructs assessed). The suggested optimal cut-off points for measures and respective indices of diagnostic utility are reported and interpreted in accordance with available recommendations for facilitating the interpretability of medical statistics reporting outcomes (Gigerenzer, 2002). CONCLUSIONS: Ability of measures to

predict diagnostic status for specific disorders was widely discrepant overall. Most index tests designed to measure symptoms associated with a specific disorder demonstrated inferior performance (relative to other non-specialized index tests) when predicting their respective diagnostic target. The practical implications of this study's results are discussed, highlighting the utility of diagnostic validity examinations employing the methodology used in the current study (Kraemer, 1992; 2013).

DEDICATION

This dissertation is dedicated to my friends, family, and my wonderful advisor, Dr. John Young.

ACKNOWLEDGEMENTS

I could not have completed this dissertation without a wonderfully supportive department and committee. Dr. Gross, Dr. Young, Dr. Bentley, and Dr. Maack have all been tremendously helpful and supportive throughout this process. I appreciate all of your help.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

1. Diagnosis of Psychopathology

Since the formation of a nosological system for psychological disorders there have been numerous deletions, additions, and revisions of diagnostic categories and criteria (Regier et al., 1998; Spitzer & First, 2005). The first two iterations of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-I and DSM-II; American Psychiatric Association, 1952, 1968) were heavily influenced by psychoanalytic theory, the ideological zeitgeist at that time (Hyman, 2003; Garb, 2003). Consistent with psychoanalytic theoretical conception of mental illness, psychopathology was defined as phenomena that were primarily psychological in nature and placed little emphasis on organic or biological causes, which required clinicians' subjective assessment of diagnostically relevant, yet unobservable, internal states (Miller, Dasher, & Collins, 2001). The iteration of the DSM-III (American Psychiatric Association, 1980) represented the result of a paradigmatic shift toward a more empirically-based approach to assessment and asserted the importance of employing objective diagnostic procedures (Spitzer & Williams, 1994; Miller et al., 2001; Bogenschutz & Nurnberg, 2000). Clinical evaluations began to be more likely to be semi-structured in nature and instead of primarily comprising open-ended questions, most items began to focus on discrete, observable symptoms and behaviors (e.g., duration of symptoms, observable behaviors, functional impairment, etc.; Miller et al., 2001; Regier et al., 1998). One major goal of the changes made in the DSM-III diagnostic criteria was to facilitate the ability for independent evaluators to arrive at the same diagnostic conclusions for

the same patient (i.e., increase inter-rater reliability; Garb, 2003). The DSM-III was a vast improvement compared to its predecessors, however, in the absence of standardized research-based assessment the issue of low levels of inter-rater reliability has remained problematic to varying degrees for every iteration of the DSM thus far (American Psychological Association, 1994; Zimmerman & Mattia, 1999; Regier et al., 1998; O'Boyle & Self, 1990; Blais & Norman, 1997; Finn, 1982).

Another complaint shared by many researchers and clinicians is that the dichotomous nature of diagnostic categorization is an oversimplification of a patient's presentation (Swets, Dawes, & Manahan, 2000; Pincus et al., 1999; Antony et al., 1994). It has been argued that while these "present" or "absent" labels are appropriate at times for some types of diagnostic decisions (e.g., when diagnosing organically-based diseases such as HIV) they fail to provide all the necessary information concerning psychological disorders because "yes" or "no" (i.e., assigning categorical cut-offs) fails to exhaust the list of alternatives (i.e., inherently dimensional phenomena; Di Nardo, Moras, Barlow, Rapee, & Brown, 1993; Antony et al., 1994; Crowe, 2000; Widiger & Samuel, 2005).

Providing an exhaustive list of DSM criticisms is beyond the scope of this brief review, however, they are numerous and it is clear that the current system is not without its limitations (see Phillips, First, & Pincus [2003] for a comprehensive review). Regardless of whether this nosological system is or is not the most appropriate for diagnostic purposes, it offers certain advantages and its usage is unavoidable in the context of a managed-care environment (i.e., reimbursement for services, meeting record-keeping requirements, etc.; Cashel, 2002; Miller & Luft, 1997; Eisman et al., 2000; Basco, 2003). Additionally, the treatments that have been identified as meeting the most stringent scientific standards (i.e., "efficacious" treatments

identified by APA Division 12 Task Force; Chambless & Hollon, 1996; 1998) are diagnosis-specific, again requiring application of a diagnostic label (Shear, Greeno, & Kang, 2000; Miller, 2001; Basco, 2003).

The selection of a particular diagnostic label, then, is an important process. In addition to determining relevant, scientifically supported treatments diagnoses are also used to make other important clinical decisions pertinent to recovery speed such as choice of medication and length of hospital stay (Garvey & Tuason, 1980; Miller et al., 2001). In some cases, especially in larger treatment facilities, a patient's treatment plan may be informed solely by the diagnosis listed in his or her chart (Basco et al., 2000). Additionally, in courtroom settings diagnostic labels applied by "experts" using questionable or demonstrably invalid assessment methods (Lilienfeld, Ruscio, & Lynn, 2008) could inform life-altering decisions (e.g., child custody, ability to stand trial, sentence type and duration, etc.; Emory, Otto, & Donohue, 2005).

In short, there is a lot riding on clinicians' diagnostic decisions. Given the potential negative consequences of assigning inaccurate diagnoses research efforts have been made to determine the degree to which diagnoses rendered in routine mental healthcare settings are accurate (Basco et al. 2000; 2002; Garb, 2003; Miller at al., 2001). A review of these studies, culminating in a rationale for the current examination, follows.

2. Diagnostic Accuracy

Before beginning a review of this literature, delineation must be made between the composition of clinical assessments conducted in practice-oriented settings (e.g., primary care, psychiatrists, outpatient mental healthcare, etc.) and those that we are using for comparison (i.e., reference criterion). Almost universally the assessment of patients' problems in routine practice

settings is conducted in an unstructured, unstandardized manner, and this interview is the basis for the diagnosis assigned by the clinician. Besides a few widely publicized, notable exceptions (e.g., Epidemiological Catchment Studies; Robins & Regier, 1991; Kessler, et al., 1994; 2005) results of epidemiological surveys have largely been based on unstructured clinical evaluations (Koenigsberg, Kaplan, Gilmore, & Cooper, 1985; Mezzich, Fabrega, Coffman, & Haley, 1989; Oldhamn & Skodol, 1991) and widely discrepant findings concerning prevalence rates are not atypical (Regier et al., 1998). In contrast, the reference criterion used for comparison with diagnoses made in clinical practice are research-based structured interviews (e.g., ADIS-IV, SCID-I, etc.), which provide comprehensive symptom coverage and use a fixed format for question wording, order, and scoring procedures. Structured clinical interviews are considered to be the "gold standard" of psychiatric diagnosis (Summerfeldt & Antony, 2002; Wetzler & van Praag, 1989; Miller, 2001; Shear et al., 2000). Accordingly, many of the nosological criticisms outlined above are diminished or eliminated when research-based structured interviews are used (i.e., diagnostic reliability; Garb, 1998; Swets, 1988).

2a. Diagnostic Accuracy in Clinical Practice

Historically much of the published research addressing issues with assessment as it is conducted in practice has been anecdotal and commentary in nature (Piotrowski, 1999). Only recently has the diagnostic process and accuracy of diagnostic interviews conducted in practice been questioned systematically (Stout, & Cook, 1999; Shear et al., 2000; Basco, 2003). Ubiquitously, the results have demonstrated unacceptably low levels of diagnostic agreement across unstructured and standardized, research-based interview procedures (for reviews see: Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009; Mitchell, Vaze, & Rao, 2009; Cepoiu et

al., 2008). Additionally, unstructured diagnostic interviews identify co-morbid conditions far less frequently than structured interview procedures (Zimmerman, 2003). Findings of low levels of inter-rater reliability when using unstructured interviews (i.e., agreement across diagnosticians in practice) suggest that the actual diagnostic process (i.e., composition of the clinical interview) varies considerably across clinicians and settings (Garb, 1989; Basco, 2003). This variability makes it very difficult to operationalize the actual assessment practices that comprise clinical interview procedures employed in practice (Addis, 2002; Mellor-Clark, Barkham, Connell, & Evans, 1999); however, diagnostic discordance is, in general, attributable to the absence of standardization in diagnostic procedures.

As would be expected, the absence of any imposed interview structure (i.e., question content, order, etc.) guiding administration leads to inadequate data collection (Zimmerman, 2003). Examinations of content coverage in routine clinical interviews show that certain symptoms are never queried, especially those that may be socially undesirable or embarrassing (Beck, Epstein, Brown, & Steer, 1988; Bagley & Genius, 1991). In a study comparing diagnostic accuracy of a traditional clinical interview format to that of a computer-administered structured interview (CADI) clinicians only gathered 33% of key diagnostic data when using the unstructured clinical interview while the (CADI) gathered 100%. Additionally, 54% of diagnoses based on the traditional interview were discrepant with the reference diagnosis (Miller, 2001). Given that clinicians have difficulty gathering enough information to make an accurate primary diagnosis, the findings concerning under-identification of co-morbid diagnoses in practice settings are not surprising (Grove & Meehl, 1996; Meehl, 2001; Frances, Widiger, & Fyer, 1990; Blashfield, 1990).

Across the literature the most frequently cited problem with assessment methods employed in practice is clinicians' reliance on "clinical judgment" for making important diagnostic and treatment planning decisions rather than data (e.g., standardized assessment protocol; Dawes, 2005; Miller, 2001). Decades of empirical research have demonstrated the fallibility of clinical judgment and the superiority of actuarial (i.e. data-driven) methods of decision-making (Grove, Andreason, McDonald-Scott, Kelle, & Shapiro, 1981; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meyer et al., 2001; Garb, 1994, 1996; 2003; Hammond, 1996; Mojtabai & Nicholson, 1995; Meehl, 1954; Luft, 1950; Dawes, Faust, & Meehl, 1989; 1993).

In terms of difficulties with clinical judgment decisions, research has identified the application of heuristics (e.g., availability, representativeness, illusory correlations, belief perseverance, etc.) and clinician biases as being often responsible for misdiagnosis (Basco, 2003; Lilienfeld, et al., 2003; Garb, 2005). Many studies have assessed the effect of these biases in diagnostic practices and repeatedly demonstrated the tendency for clinicians to seek confirming information when they have a "hunch" about a client's personality or potential symptom presentation (e.g., when clinicians are given false information a priori; Myers et al., 2002). Clinicians' preconceptions about the nonclinical features of the patient (e.g., age, race, sex, etc.) cause them to focus on some issues more and neglect others (Norman & Brooks, 1997). Numerous studies have demonstrated that varying personal characteristics (e.g., race, age, gender, social class) of hypothetical patients described in clinical vignettes results in clinicians assigning different diagnostic labels despite identical symptom presentation (Garb, 1995; 1997; Van Ryn & Burke, 2000; Loring & Powell, 1988; Widiger & Spitzer, 1991; Becker & Lamb, 1994). The setting in which the diagnostic interview takes place (e.g., juvenile justice center) has also been shown to influence the areas queried leading to misdiagnosis (Kendall-Taylor &

Mikulak, 2009). It would seem that diagnostic accuracy would be improved by having face-to-face contact with a patient; however, meta-analytic results suggest that clinicians are *less* accurate when assigning diagnoses based on a clinical interview as opposed to simply reviewing a file (Loy & Irwig, 2004; Myers at al., 2002). Even aspects of a patient's speech and affect presentation during the clinical interview have been shown to influence diagnostic assignment, which means that the same patient could, hypothetically, receive different diagnoses on different days (Grigg, 1958; Garb, 2003).

One might assume, and many clinicians assert, that clinical experience improves clinicians' ability to accurately identify psychological disorders; however, this is simply not true (Myers at al., 2002; Meyer at al., 2001; Christensen & Jacobson, 1994; Brammer, 2002; Brehmer, 1980; Brenner & Howard, 1976). When the performance of experienced clinicians has been compared to that of less-experienced clinicians, the "expert" clinicians have reliably performed no better than comparison groups (Garb, 1989; Turner, 1966; Graham, 1967; Silverman, 1969; Dawes, 1994). This outcome has been reproduced across multiple conditions (e.g., interview data, observation data, biographical information, projective protocols, and personality test protocols) when comparing the accuracy of experienced clinicians' ratings of personality and psychopathology to those of less-experienced individuals (e.g., graduate and undergraduate students, masters level clinicians; Garb, 1989; Anthony, 1968; Grigg, 1958; Faust et al., 1988; Oskamp, 1962; 1965; Brammer, 2002; Gadol, 1969; Turner, 1966; Clark, Watson, & Reynolds, 1995). Additionally, studies assessing the relationship between clinicians' level of confidence in their diagnostic accuracy and their actual accuracy have failed to demonstrate a significant correlation between the two (Goldberg, 1959; 1968; Grove, 1987; Bickman, 1999; Garb. 2003).

Finally, diagnosis based solely on a clinical interview promotes the monomethod/monooperation bias (Cook & Campbell, 1979), which can be summarized to state that one data source is insufficient for basing diagnostic decisions and validity is maximized when variables of interest are assessed using multiple measures. Researchers have demonstrated empirically that the quality of clinical assessment in the practice setting is improved by integrating data obtained using multiple methods of assessment (Basco et al., 2000; Meyer et al., 2001). In a routine practice setting, Basco et al. (2004) found that the addition of medical records to a structured interview (i.e., Structured Clinical Interview for DSM-IV Axis I Disorders; *SCID-I*) resulted in a higher rate of agreement (kappa coefficient .76) with the reference criterion (i.e., independently verified diagnosis by psychologist) than the SCID-I alone (kappa coefficient of .61) for both primary and secondary disorders. The addition of supplemental information has demonstrated improved diagnostic accuracy in a diverse range of settings and diagnoses including adding an informant report to cognitive testing in dementia identification (Mackinnon & Mulligan, 1998) and supplementing medical test results with clinical information (Loy & Irwig, 2004).

2b. Structured Diagnostic Interviews

Research across multiple disciplines conducted since the publication of Meehl's seminal text *Statistical Prediction* (1956) comparing data-driven and intuitive decision-making approaches has overwhelmingly been in the favor of data-driven methods (to such a degree that the statistical versus clinical prediction "debate" is now a misnomer; Dawes, Faust, & Meehl, 1993). Specifically related to psychodiagnostic approaches, research assessing accuracy of available data-driven diagnostic assessment has concluded that structured and semi-structured

interview procedures (e.g., ADIS-IV, SCID-II, etc.) are psychometrically superior to unstructured (i.e., intuitive) assessments (Summerfeldt & Antony, 2002; Garb, 1998; Rogers, Gordon, Schanzenbacher, & Casey-Smith, 2001). However, despite these procedures being considered the field's "gold standard" diagnostic procedures they are rarely employed in practice-oriented environments (Piotrowski, 1992a; 1992b; 1999; Piotrowski, Belter, & Keller, 1998; Phelps, Eisman, & Kohout, 1998; Zimmerman, 2003).

Ideally clinicians would adhere to the highest possible standard of care through frequent implementation of the strongest, most consistent assessment techniques; however, the reality is that extant barriers in practice-oriented environments often make application of this diagnostic approach infeasible at best and impossible at worst (Chorpita & Nakamura, 1998; Miller & Luft, 1997; Eisman et al., 2000). Many of the barriers to research-based assessment practices are identical to those that impede provision of scientifically supported interventions in these settings (i.e., the "science-practice gap"; Addis, 2002; Addis, M. E., & Krasnow, 2000; Zimmerman, 2003; Weisz & Addis, 2006; Wood, Garb, Lilienfeld, & Nezworski, 2002).

As a result of evolving managed-care organizations and expansion of efforts to obtain reimbursement for services through third-party payers, providers have been forced to truncate a wide range of psychological services (Eisman et al., 2000; Acklin, 1996). Clinical assessment practices have been hit especially hard with managed-care directives severely curtailing or flat out denying coverage for psychodiagnostic assessment (Piotrowski, 1999; Christenson & Jacobson, 1994; Butcher, 1997; Miller, 1999). In surveys assessing changes in assessment practices as a result of managed care directives, practitioners have reported a decline in the usage of lengthy or costly assessments (Addis, M. E., & Krasnow, 2002; Cashel, 2002; Piotrowski, Belter, & Keller, 1998). On the whole, third-party payers have criticized psychological

assessment practices as having little utility in the context of symptom-improvement to justify eliminating reimbursement for assessment activities (Eisman et al., 2000; Piotrowski, 1999). Griffith reported that in a survey of nine managed care organizations the consensus was that psychodiagnostic testing did not have diagnostic or clinical utility (Griffith, 1997). Providers suggested that information gathered through clinical interviews is adequate for diagnostic purposes and justified this stance by referencing the absence of any mention of psychological assessment in the DSM-IV (APA, 1994; Griffith, 1997). Admittedly, it is somewhat understandable how mental-healthcare stakeholders who are only vaguely (if at all) familiar with issues related to psychological treatment would question the utility of assessment. The relationship between diagnostic assessment and treatment-related variables is not intuitive; however, dissemination of information concerning the treatment implications of accurate diagnoses is the responsibility of researchers and promoters of evidence-based psychological practice (Myers at al., 2002; Piotrowski, 1999).

The most frequently cited obstacle to empirically-based diagnostic assessment is time constraint, which is largely a consequence of managed-care directives limiting amount of time that is reimbursable (Piotrowski, 1999; Addis, 2002; Bickman, 1999; Glasgow, Lichtenstein, & Marcus, 2003). Providing thorough diagnostic assessments is time-consuming and is therefore incompatible with most managed care organizations' primary goal of cost-containment (Piotrowski & Keller, 1992; Zimmerman, 2003; Basco et al., 2000). Clinicians are typically expected to provide diagnoses following the first session (i.e., the intake) in order to meet requirements imposed by third-party payers and other various stakeholders in the provision of mental health treatment (Miller & Luft, 1997; Piotrowski, 1999). If this were the only information to be gleaned from this initial session then perhaps the administration of a structured

or semi-structured assessment procedure would be feasible; however, this is unfortunately not the case. In most managed care settings the initial session is used to collect all pertinent information for treatment and to meet record-keeping requirements (e.g., insurance information, history, biographical data, etc.; Piotrowski et al., 1998). The amount of time the clinician is allowed to complete the diagnostic assessment varies significantly depending upon setting; however, researchers have cited clinicians' reports of as little as 30 to 45 minutes allotted to complete both the evaluation and write-up (Zimmerman, 2003; Basco et al., 2000). Clinicians are struggling to conduct diagnostic evaluations more efficiently in the face of declining reimbursement rates; however, the administration of a structured interview that can take upwards of two hours is simply infeasible (Addis, 2002).

As a result, time and cost-efficiency often take precedence over diagnostic accuracy, which is why we see practitioners using less reliable, unstandardized assessment procedures (Hatfield & Ogles, 2007; Piotrowski, 1999). Unfortunately, it seems that economic reality, rather than clinical utility, now guides psychodiagnostic practices and test selection (Zimmerman, 2003). Therefore the limitations that exist in "real-world" practice settings must become a consideration in assessment research if it is to be translated to practice (Piotrowski, 1999; Yates & Taub, 2003; Jensen Doss, 2005; Bickman, 1999).

3. Diagnostic Screening Research

Clearly the way clinical assessment is conducted must change if it is to both make the cuts imposed by managed care (Wood, Widiger, & Sankis, 2000) and simultaneously be informed by scientific findings. Assessment strategies must be cost-effective in addition to demonstrating tangible benefits and incremental validity over those currently employed

(Pitrowski, 1999). Administration of the most effective diagnostic tools is time-consuming and therefore incompatible with cost-containment. Fortunately, while structured and semi-structured interviews may be the *optimal* diagnostic tool, they are not the o*nly* tool at our disposal (Murphy et al., 1987).

A promising practical approach to enhancing diagnostic accuracy within the clinical setting is the usage of diagnostic screening devices prior to interview administration (e.g., Zimmerman & Mattia, 2001b; Chorpita, Bernstein, & Daleiden, 2008; Chorpita & Nakamura, 1998; Kaufman et al. 1997; Lucas et al. 2001). In the last decade there have been hundreds of self-report devices developed including some designed specifically to screen for Axis-I disorders (Foa, Riggs, Dancu, & Rothbaum, 1993; Thelen et al., 1991; Zimmerman, 2003; Zimmerman & Sheeran, 2003Weisz & Addis, 2006; Yates & Taub, 2003). The diagnostic utility (i.e., level of agreement between a questionnaire and a diagnostic interview) of these measures obviously varies across instruments; however, results have demonstrated that screening devices' agreement with the gold standard can be as high as the test-retest reliability of the diagnostic interview itself (Zimmerman & Coryell, 1988).

Developing or identifying existing screening instruments whose performance is close to that of the gold standard would drastically reduce the amount of time needed for assessment or, as the case may be in clinical practice settings, increase the likelihood that *any* form of actuarial prediction will be used. Screeners cannot provide definitive diagnoses, although they can guide the amount of focus given to various to symptom domains by providing predictive information a priori. It is apparent how this actuarial approach so strongly advocated by Meehl (1956) could prevent many of the problems arising from clinician bias.

This approach is based on the understanding that within a clinical context the cost-benefit ratio of an assessment strategy can only be determined on an individual patient basis meaning that certain tests will have more benefits for some patients than others (Cronbach & Gleser, 1965; Finn, 1982). For example, the social phobia module of the ADIS-IV may have excellent cost-benefit ratio for one person (e.g., someone with social phobia symptoms) while a very low cost-benefit ratio for another person (e.g., someone exhibiting only externalizing symptoms; Meyer, et al. 2001). Likewise, the administration of a test that is too long or costly for everyday use may be essential for obtaining a complete diagnostic picture for a particular patient (e.g., extensive memory battery with an elderly patient). For this reason Meyer et al. (2001) suggests that the traditional practice of using a standard assessment protocol with every patient is more costly and provides a less accurate diagnostic profile (i.e., due to administering unnecessary tests; Griffith, 1997). Being able to make this determination prior to conducting the diagnostic interview greatly increases diagnostic efficiency while not compromising accuracy.

Several studies exist demonstrating the utility and feasibility of self-report measures for use as diagnostic screening tools in routine clinical practice settings. Zimmerman & Mattia (1999c; 2001a; 2001b) developed a brief self-report scale Psychiatric Diagnostic Screening Questionnaire (PDSQ) designed to assess for the most common Axis I disorders and examined its diagnostic utility within two independent samples (630 outpatients in a pilot study and 670 outpatients in a replication). The PDSQ demonstrated excellent sensitivity (i.e., probability that a condition will be detected) and negative predictive power (i.e., probability that a negative condition will be correctly identified) for identification of both primary and secondary (i.e., comorbid) disorders within each sample. Chorpita and Nakamura (1998) employed a more complex procedure in which they compared diagnostic reliability of scoring based on the

standard administration of the ADIS-C/P to that of a modified administration procedure informed by self-report scores using an algorithm based on Bayesian logic. When participants' diagnoses were compared based on the two different scoring procedures (as calculated by standard and dynamic algorithms) the results demonstrated that scoring using the dynamic interview algorithm reduced administration time while obtaining equal or improved diagnostic accuracy. The authors indicate that the positive results obtained in their study have implications for practice environments even in the absence of integrating complex predictive algorithms.

Not only did these studies demonstrate that screening tools "worked" (i.e., accurately predicted diagnostic classification) in clinical populations, they also demonstrated the feasibility of implementing screenings in routine practice settings. A high level of agreement between self-reports and semi-structured interviews does little good if we cannot implement necessary procedures in actual clinical practice (Meyer et al., 2001). Research concerning "efficacy versus effectiveness" has demonstrated that results from tightly controlled treatment studies (i.e., randomized controlled trials) do not necessarily translate to real-world practice (Glasgow et al., 2003). One of the primary concerns is the stringent inclusion and exclusion criteria that is often applied in efficacy trials, which was addressed in Zimmerman and Mattia's study by using outpatients presenting to "real-world" clinical settings. Issues often arise due to incompatibility with existing organizational structure and demands and, as mentioned previously, the feasibility of administering semi-structured interviews is severely limited due to this problem (Piotrowski, 1999; Meyer et al., 2001). Alternatively, the usage of self-report measures as diagnostic screening devices offers the advantage of being efficient due to the administration method in terms of enhancing feasibility (i.e., they can be completed in the by the patient in the waiting room; Zimmerman, 2008). Further, they are compatible with the existing organizational structure

in many ways given that self-report questionnaires are already commonly used in an array of medical settings to obtain patient histories prior to initial evaluation (Piotrowski, 1999).

Surveys of clinicians' assessment practices further support the feasibility of self-report administration in the context of a managed-care environment. Cashel (2002) asked a sample of clinicians (n = 162) how their assessment practices have changed as a result of managed-care related changes (i.e., more or less usage of specific youth measures). Out of the 45 youth-focused measures reported by respondents the most negatively affected measures in terms of continued implementation were the Wechsler Intelligence Scale for Children (WISC): 22.8% less usage, Rorschach: 16.7% less usage, and the Minnesota Multiphasic Personality Inventory **(**MMPI): 11.7% less usage. The category of measures that was least affected in terms of decreased usage, and in fact was reported to have increased in usage in some cases, were standardized, well-researched, self-report measures (e.g., Child Behavior Checklist [CBCL], Connor's Teacher and Parent Rating Scales, Children's Depression Inventory [CDI], Teacher Report Form [TRF], etc.). No information was gathered concerning the reason for these changes, however, the authors speculated that the most likely variables responsible were time and cost given that usage of the most time intensive and costly measures decreased while the shorter, less costly measures' usage increased. Administration and scoring are quick and normative information is generally available for empirically-based instruments, which allows the clinician to draw clinical inferences upon scoring patients' responses. In fact, some self-report measures have freely available scoring programs that are accessible online (e.g., Revised Child Anxiety and Depression Scale; Chorpita, Yim, Moffitt, Umemoto, & Francis, 2000; Trent et al., 2012)

4. Summary and Techniques for Additional Study

The accumulated evidence reporting diagnostic discordance when using gold-standard reference criterion or unstructured interviews in addition to studies revealing low diagnostic inter-rater reliability strongly support the notion that globally different assessment methods should be used in routine clinical practice (Lucas et al, 2001). This problem has received more attention over the last decade in the face of declining reimbursement rates by third-party payers, which has often meant cutting funding for assessment (Zimmerman, 2003; Zimmerman et al., 2007; Garb, 1998). In this context, which is not likely to substantially change in the foreseeable future, it is clear that more efficient assessment practices must be identified if actuarial prediction methods are going to be implemented (Zimmerman, 2003; Swets et al., 2000). An alternative to lengthy structured interviews is the usage of self-reports, which have empirically demonstrated the ability to provide a practical, efficient and accurate method of assessing patients' symptoms (Zimmerman, 2003; Garb, 1998). Perhaps managed healthcare would be more likely to reimburse assessment practices if research demonstrated the potential for effective diagnostic practices and outcome monitoring to cut healthcare costs in the long-run (e.g., providing diagnosis-specific treatment, terminating therapy when symptoms abate, changing strategy if no change is observed, etc.). Obviously, diagnostic screening procedures do not take the place of structured interviews; however, the application of these assures that at least *some* science is being implemented within practice settings and evidence from the studies reviewed above suggests that their integration into typical clinical strategies would enhance accurate assessment.

4a. Identifying valid measures

Accrued research suggests that self-report instruments are a practical, feasible option in frontline care settings and provide valid incremental information for diagnostic purposes. However, it is important to acknowledge that not all measures are created equal; that is to say, they possess varying levels of psychometric support necessary for assumptions concerning diagnostic validity or relationship to the construct of interest (Deeks, 2001). Practice settings often generate their own unstandardized self-report assessments, which are not appropriate for use as diagnostic screening tools (Zimmerman, 2009). In contrast to empirically-derived measures that possess normative guidelines for interpretation and established psychometric properties; intuitively-derived measures lack the information necessary to guide clinicians' interpretation of respondents' results (Zimmerman, 2003).

When assessing the diagnostic utility of self-report instruments as screening devices, the primary outcome is the instrument's predictive validity. If the outcomes of screening tools do not predict group membership, with a reasonable and quantifiable amount of certainty, then their administration is a waste of (precious) time (Meyer et al., 2001). The most common way to assess an instrument's diagnostic utility (e.g., ability to accurately classify those who have the disorder from those who do not and minimize inaccurate classifications) is to use a receiver operating characteristics (ROC) curve (Tripepi, Jager, Dekkar, & Zoccali, 2009).

4b. Receiver and Operating Characteristics (ROC) Curve Analysis

Receiver operating characteristics (ROC) curve analyses answer questions concerning the decision-making ability of a measure. Statistical tests using the receiver operating characteristic (ROC) curve were first used to determine signal detection ability in psychophysics research (Green & Swets, 1966) and have since been applied to decision-making questions across

multiple disciplines (Brown & Davis, 2006). Information provided by ROC analyses has been especially useful in the medical field given the often costly and/or invasive nature of "gold-standard" diagnostic tests. ROC analyses are often used to assess the ability of a shorter, more feasible diagnostic tests to be used for screening purposes (Pereira-Maxwell, 1998). The application of a screening device entails dichotomous decision-making using continuous data, which is not atypical in medical settings (Tripepi et al., 2009) and ROC analyses are used to determine the optimal cut-off to dichotomize a continuous scale (i.e., score that best predicts membership in "positive" or "negative" group for a given disorder/disease). Application of the "gold standard" assessment method for the disorder/disease in question dictates *actual* group membership and is the metric for comparison (Streiner & Cairney, 2007). In psychological research this is usually a form of a research-based structured interview. Albeit imperfect, structured interview procedures are the most valid measurement tool currently available in psychiatric diagnosis (Tripepi et al., 2009), and thus have provided the majority of comparison data for research in this area.

The standard statistical indices of diagnostic utility reported are sensitivity, specificity, positive predictive power, and negative predictive power. Sensitivity is the test's ability to identify positive cases (also called true positive rate; percent of ill persons who are identified by the test as ill). Specificity is the test's ability to identify negative cases (also called true negative rate; percentage of non-ill persons correctly identified by the test as non-ill). The positive and negative predictive values are generally considered more clinically meaningful than sensitivity and specificity values (Kessel & Zimmerman, 1993). These values indicate the probability that an individual is ill or non-ill given that the test identifies him or her as ill or non-ill. So positive predictive value is the percentage of individuals classified by a test as ill who are truly ill, while

negative predictive value is the percentage of individuals classified as non-ill by the test who are truly non-ill (Linden, 2006).

In order to identify optimal cut-off scores for maximizing accurate prediction ROC analyses examine sensitivity and specificity across the entire range of cut-off values, which are plotted against each other on a scatterplot graph to produce the ROC curve (Tripepi et al., 2009). Calculating the area under the ROC curve (Area Under the Curve, or AUC) provides an index of the overall discrimination performance of the test and can be considered a global estimate of diagnostic accuracy (Streiner & Norman, 2003). The area that falls under the ROC curve is quantified (i.e., Area Under the ROC Curve; AUC) and can range from .5 (no discrimination; diagonal line across graph) to 1 (perfect discrimination; curve would comprise a vertical line to the top off the Y-axis then curve sharply into a line straight horizontal line that continues to the right-most boundary of the graph). Generally guidelines dictate that the accuracy of tests with AUC values falling between .5 and .7 are considered low, .7 and .9 moderate, and above .9 high (Fischer, Bachmann, & Jaeschke, 2003). The AUC can be interpreted as the percentage of time that a positive case will have a higher score on the test than a negative case (Hanley & McNeil, 1982). Points (representing scores) plotted closer to the upper left corner of the ROC curve are associated with the highest accuracy (i.e., optimal balance between sensitivity and specificity; Tripepi et al., 2009), however, this score may or may not be the most desirable cut-off score depending on the purpose of the test (Streiner & Cairney, 2007).

Additionally, Kessel & Zimmerman (1993) suggest reporting Kappa (a statistical metric of reliability that corrects for chance agreement). Kappa represents the level of agreement between the test in question and a gold-standard beyond that accounted for by chance alone. Although there are a variety of statistics available that can be used to account for chance

agreement, kappa is one of the most widely recognized in psychological research. The overall classification rate (also know as the “hit rate” or “overall level of agreement”) refers to the proportion of ill and non-ill patients correctly classified by the test (McFall & Treat, 1999).

Diagnostic validity studies often vary considerably regarding which statistical indices are reported; however, different outcomes provide different perspectives of performance and can be used in combination to establish a broad profile of the utility of diagnostic tests (Kessel & Zimmerman, 1993). More or less emphasis may be placed on one statistic over another in accord with the nature of the test, the population tested, and the hypothesis of the investigation (Basco et al., 2000). Determination of the optimal cut-off score, or threshold for inclusion, is based on the aforementioned statistical outcomes in conjunction with considerations regarding the purpose of the measure and “costs” of type I (i.e., false positive) and type II errors (i.e., false negative; Linden, 2006; Metz, 1978). Dependent upon these variables more or less weight can be assigned to the likelihood of obtaining a false positive or false negative. Within a mental-health setting the costs associated with inaccurately assigning a positive label to an individual are minimal, but these would likely be clinician time and potentially discouraging usage of screeners (e.g., clinicians may doubt efficacy of screener if it frequently produces false positives suggesting usage of more in-depth assessment procedures that are unnecessary). However, the potential costs associated with failing to identify a positive case are more consequential including decreased likelihood that a disordered individual will receive any treatment or be accurately diagnosed in the future (i.e., leading to persistence and/or worsening symptoms). Therefore, when using a measure for diagnostic screening purposes the cut-off for case identification often errs on the side of being over-inclusive rather than under-inclusive (i.e., greater chance of type one errors than type two; Zimmerman, 2008) by setting the threshold low producing good

diagnostic sensitivity (i.e., unlikely to miss a positive case) and high negative predictive power (i.e., those who fall below the threshold are unlikely to be a positive case). This means non-disordered individuals may screen positively and be classified inaccurately, but disordered individuals are not likely to be falsely labeled and thus overlooked (Zimmerman, 2009; Streiner & Cairney, 2007; Streiner & Norman, 2003; 2006).

5. Current Study

Informed by previous research on structured interviewing, diagnostic reliability, signal detection, and the need to enhance accurate assessment in clinical practice this study examined the diagnostic utility of a wide cross-section of freely-available self-report instruments in terms of predicting psychiatric diagnoses. The gold-standard determination of diagnostic outcome was a research-based semi structured interview; specifically, the Anxiety Disorders Interview Schedule for the DSM-IV (ADIS-IV) and a module of the Structured Clinical Interview for the DSM-IV Axis I Disorders (SCID-I). The battery of measures was subjected to multiple sets of Receiver Operating Characteristics (ROC) analyses predicting multiple levels of diagnostic classification derived from the reference criterion: presence of any psychopathology (i.e., ≥ 1 disorder), presence of comorbid disorders (i.e., ≥ 2 disorders), and participants' diagnostic status on each individual diagnosis (i.e., specific anxiety disorders, major depressive disorder, and eating disorders – described in further detail in the METHODS section below). Descriptive data and diagnostic validity outcomes will be reported for all of the aforementioned diagnostic categories that best capture the important facets of diagnostic performance guided by the most recent methodological literature in this domain (i.e., Pepe, 2003; Kraemer, 2003). Additionally,

the results will be communicated using format and terminology intended to facilitate the ease of interpretation and clinical application (as suggested by Gigerenzer, 2002).

In more general terms, the current study aims to answer the following research questions: a) What is the correspondence between self-report measures of psychopathology intended to measure general distress and symptomology of specific diagnoses in a non-clinical sample of young adults using a paired-group design?; b) How well does each of the measures predict a diagnostic classification based on the "gold-standard" reference criterion (i.e., ADIS-IV & eating disorders module of the SCID-I); and c) Can these briefer measures be used as screening devices that could then be followed up with more time-intensive assessment strategies in a way that minimizes clinical assessment burden in applied settings?

II. METHODS

1. Participants

The current study utilized an archival dataset comprising data collected from a sample of undergraduate students attending a large, public university in the south. Age and ethnic composition of the sample can be found in Table 1[1]. The University of Mississippi Institutional Review Board approved all procedures.

2. Procedure

Data used in the current investigation came from a large-scale research project examining the interaction of disgust and a variety of psychological constructs. Undergraduate students enrolled in introductory psychology classes participated in this study in order to obtain course credit. The study included administration of semi-structured diagnostic interviews (described in more detail below) and a self-report battery comprising a variety of measures intended to measure an array of psychological phenomena.

Given the long-term nature of the project, and changing focus of the research study over time, additional outcome measures were added to the screening battery partway through the larger project's data collection process. Thus, 94 participants from the current sample completed an initial version of the screening battery. Another 198 participants completed a second version of the battery, which included the same measures as the first as well as the following: the

[1] All Tables are located in APPENDIX I of this document.

Acceptance and Action Questionnaire (AAQ), Fear Survey Schedule III (FSS), PTSD Symptom Checklist (PCL), Penn State Worry Questionnaire (PSWQ), Social Interaction Anxiety Scale (SIAS), Social Phobia Scale (SPS), and the White Bear Suppression Inventory (WBSI). Each measure is described in greater detail below.

Diagnoses used in the current study were assigned based on an interview conducted using the Anxiety Disorders Interview Schedule for the DSM-IV (ADIS-IV; anxiety and depressive disorders) and the eating disorders module of the Structured Clinical Interview for the DSM-IV Axis I Disorders (SCID-I). Diagnoses derived from these interviews used in the current study were major depressive disorder, generalized anxiety disorder, obsessive-compulsive disorder, panic disorder, social anxiety disorder, specific phobia, post-traumatic stress disorder, and eating disorders (i.e., anorexia nervosa, bulimia nervosa, and not otherwise specified[1]). Interviewers were research assistants working on a Bachelor's degree in psychology who were provided training by a senior-level graduate student, who in turn was trained by a supervising doctoral-level licensed clinical psychologist. Prior to assisting in the research study's data collection process, research assistants participated in a mock interview with the supervising doctoral-level psychologist to assure interviewing competency and obtain approval. In addition to this initial screening process, periodic updates in training were provided to limit potential drift in interview implementation fidelity. Recombination of the symptom-level data produced from semi-structured diagnostic interviews and subsequent determination of which individuals fulfilled criteria for a psychiatric diagnosis was made by one of the licensed, Ph.D.-level principal investigators.

[1] Due to the low prevalence of anorexia nervosa (*n*=4), bulimia nervosa (*n*=14), and eating disorder not otherwise specified (*n*=17) identified in the current sample, these diagnoses were collapsed into a single "eating disorder" category (*n*=35). This strategy also comports with the eating-disorder symptoms self-report measure, which focuses on general symptoms as opposed to diagnosis-specific behaviors.

3. Measures

3a. Diagnostic Interviews

Anxiety Disorders Interview Schedule for DSM-IV Disorders (ADIS-IV; Brown et al., 1994). The ADIS-IV is a structured interview designed to assess for current episodes of Axis I disorders (e.g., anxiety or mood disorders) and allows for differential diagnosis among the anxiety disorders using DSM-IV diagnostic criteria. Diagnoses used in the present study (i.e., generalized anxiety disorder, obsessive compulsive disorder, panic disorder, post-traumatic stress disorder, social phobia, specific phobia, and depression) were assigned when the overall severity level of threshold symptoms was rated 4 or higher on a 0-to-8 scale for the respective disorder's module (Summerfeldt & Antony, 2002).

Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I; Spitzer, Williams, Gibbon, & First, 1992). The Eating Disorders module of the SCID-I was used to diagnose Anorexia, Bulimia, and Eating Disorders Not Otherwise Specified consistent with criteria outlined in the *DSM-IV-TR* (American Psychiatric Association, 2000). Diagnoses were assigned when overall severity for threshold symptoms was rated 4 or higher on a 0-to-8 scale. The SCID-I has been used in numerous studies of older adults (e.g., Stanley et al., 2003), and comparable psychometric properties for younger adults have been established (Segal, Hersen, Van Hasselt, Kabacoff, & Roth, 1993).

3b. Index Tests

Select self-reports from this screening battery were chosen for diagnostic prediction in the current study if they assessed symptoms of psychopathology relevant to diagnostic information obtained through the semi-structured interview procedure. This strategy resulted in the inclusion of 15 total measures that corresponded to diagnoses assessed in the structured interviews outlined above. Data at the level of specific index tests used in the current study are provided in Table 3, including: measure name, measure abbreviation, number of items, and reference to instrument development articles. Additionally, descriptive data regarding performance on selected measures within the current sample are also provided in Table 3, including mean score, standard deviation, Cronbach's alpha, number of participants used in analyses, and proportion of missing data within each scale.

Consistent with enhancing the feasibility of screening procedures the most parsimonious scoring strategies identified in previous research were used to predict diagnostic classification, which was unitary in nature for all instruments except the BIS/BAS (two-factor; behavioral inhibition and behavioral activation subscales were used for prediction) and the DASS-21 (three-factor and total score; depression, anxiety, and stress subscales were utilized as independent predictors in addition to the total score).

4. Data Analytic Strategy

4a. Data preparation.

SPSS 21.0 was used for data screening and preliminary analyses including examination of skew, kurtosis, pattern of missingness, and data imputation. Values obtained for skew and

kurtosis conformed to the criteria outlined in seminal publications guiding the interpretation of data distribution examinations (i.e., acceptable values are less than three for absolute skew and less than ten for kurtosis; Kline, 2010)[1]. All participants in the current study were administered a diagnostic interview that determined their status on the reference criterion (i.e., ADIS-IV and eating disorders module of the SCID), therefore there were no missing data on this variable. In order to maximize the usable sample size and to avoid the possible introduction of bias due to missing data, participants' data were retained and analyzed in instances that missing data comprised no more than ten percent of responses on a given scale and these values were imputed using an Expectation-Maximization (EM) algorithm approach. The EM procedure produces less biased outcomes than other strategies for handling missing data (e.g., mean substitution or listwise deletion) and is considered an acceptable alternative to the most intensive data-replacement strategies (i.e., Multiple Imputation [MI]; Shafer, 1997, 1999; Shafer & Olsen, 1998). This method utilizes multiple regression equations based on an empirically-derived estimation of the given dataset's model parameters (i.e., means and covariances) to predict missing values (Dempster, Laird, & Rubin, 1977; Enders, 2010). Assumptions for Little's MCAR test were met suggesting that the probability of missingness on any and all outcomes in the observable range of data was unrelated to any of the remaining variables in the dataset (Little & Rubin, 1987; 1988), and that the data were thus well-suited to the utilization of the EM imputation procedure overall. The EM procedure was conducted using the Missing Value Analysis module in SPSS (version 21). A thorough theoretical and statistical rationale of the EM algorithm and its solution is provided by Shafer (1997, 1999) and Shafer and Olsen (1998).

[1] According to Tabachnick and Fidell (2007; 2012) underestimation of variance due to kurtosis disappears in samples $n > 200$ participants.

Pairwise deletion of scale scores was utilized when missing data comprised more than ten percent of responses on a given scale).[1]

4b. ROC Analyses

Comparing diagnostic accuracy. The current study followed the methodology for tests of this nature (i.e., assessing diagnostic utility) outlined by Kraemer (1992), which was itself the basis for formulating the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines for reporting diagnostic characteristics of a test (Bossuyt et al., 2003). The STARD criteria facilitate comprehensive and accurate reporting of outcomes in diagnostic research and provide a checklist of issues that should be addressed when conducting research in this domain (e.g., inclusion of the term "diagnostic accuracy" in the title of the manuscript, sufficiently detailed explication of the aims of the study and/or research question, description of population characteristics relevant to the research question, detailed information regarding sampling strategy utilized, etc.; see Bossyuyt et al., 2003 for complete 25-item checklist).

The primary criterion for all analyses grouped participants into two categories: (0) those with no diagnosis for the disorder(s) of interest and (1) those with a diagnosis. Diagnostic efficiency outcome data to follow are presented for all measures grouped by the diagnostic category being predicted. These analyses were conducted using all measures for all diagnoses, independent of what constructs the measures were originally designed to capture (Streiner & Cairney, 2007). The rationale for this strategy is based on accumulated research findings suggesting that anxiety and depression overlap significantly (e.g., Brown, Chorpita, & Barlow,

[1] Number of cases removed by scale ranged from zero to three. Fear Survey Schedule-III (FSS-III) had the most cases (i.e., three) removed due to exceeding missingness threshold (i.e., >10% missing).

1998; Clark & Watson, 1991) and the development of transdiagnostic treatment protocols designed to address generalities rather that specific manifest symptoms (e.g., Allen, McHugh, & Barlow, 2008). Therefore, a measure that was not specifically developed for, or conventionally used to assess a particular facet of psychopathology (e.g., associated with a specific diagnosis) may adequately assess that construct to varying degrees, and is thus an important part of this investigation. Furthermore, information derived from this "all in" approach provides insight into the measures' discriminant validity (i.e., do measures designed to assess symptoms of a specific disorder achieve this goal for the respective disorder and not for the others?).

On the basis of the ROC analyses with weighted kappa set at .5 (i.e., equal weight assigned to false positives and false negatives) statistical indices are reported for each measure by diagnosis (Fawcett, 2004). See Table 5 for a comprehensive list and description of ROC outcomes and methods of interpretation used in the current study. The overall diagnostic efficiency of each test was quantified using nonparametric estimates of the area under the curve (AUROC) from receiver operating characteristic analyses (Kraemer, 1992) and comparisons of the diagnostic efficiency of the different index tests were calculated using the z test of dependent AUCs (Hanley and McNeil, 1983). Additionally, ROC and Q-ROC (i.e., *Quality ROC Curves)* figures were produced for visual inspection to corroborate quantitative findings (of optimal test points)[1]. The Q-ROC test disregards irrelevant scores (i.e., at which few participants score; typically at the lowest or highest range) when calculating optimal test points, thus increasing the clinical applicability of the result (e.g., test score of zero would produce one hundred percent specificity, but is not clinically useful; Kraemer, 2003).

[1]Conclusions derived from visual insepction of ROC and Q-ROC figures did not conflict with those based upon reported outcomes for diagnostic efficiency (outlined in Table 5; outcomes depicted in Tables 6 through 15). Due to the extensiveness of quantitative outcome reported in the current study, ROC and Q-ROC figures were not depicted in RESULTS section. ROC and Q-ROC figures are available from author upon request.

III. RESULTS

1. Sample Characteristics

Demographic characteristics of the sample are presented in Table 1. The majority of the sample was 18 or 19 years old (67.1%) and female (77.1%). Ethnically, the sample primarily comprised White (65.1%) and Black (28.4%) students with a small proportion self-identifying as Asian/Pacific Islander (2.7%) or Multiethnic/Other (3.4%). Prevalence rates of psychopathology in the current sample are also presented in Table 1. Overall, 66.3% of the current sample met criteria for at least one psychological disorder and 37.8% met criteria for two or more disorders (leaving 33.7% who did not meet criteria for any disorders). In terms of specific diagnoses the most common was specific phobia (*n*=121; 41.3%), while the least common was obsessive-compulsive disorder (*n*=5; 1.7%). See Table 1 for the prevalence rates of other diagnoses predicted in the current study.

2. Descriptive Data for Index Tests

Pearson correlation coefficients were calculated between the measures used in this study (See Table 2[1] for correlation matrix). As expected, significant correlations routinely emerged between measures (i.e., given overlap of symptoms and psychological constructs assessed). As can be seen in Table 2, the Behavioral Activation Scale (BAS) and the Discomfort Intolerance Scale (DIS) demonstrated the lowest levels of correlation with other measures used for

[1] See note included in Table 2 for description of sample size variability across pairs of measures.

prediction in the current study. Correlations between subscales of the same measure (e.g., DASS subscales and total score) were not at levels that would suggest they were measuring the exact same construct, supporting their potential to distinguish between diagnoses and/or provide incremental (diagnostically relevant) information.

Information at the measurement level is presented in Table 3 including scale name, abbreviation, number of items, mean score, standard deviation, internal consistency (i.e., Cronbach's alpha), sample size utilized[1], percent of data replaced (using the EM algorithm described in METHODS) for each measure used in this study, and reference to measure development article. Overall, values obtained for scales' internal consistency varied considerably, ranging from .287 on a measure developed to assess eating disorders (i.e., the Eating Attitudes Test-26) to .973 on a measure assessing specific phobia symptoms (i.e., the Fear Survey Schedule III). The majority of scales (excluding the EAT-26) obtained values for internal consistency that were ≥.7, with the Acceptance and Action Questionnaire (AAQ) and Discomfort Intolerance Scale (DIS) being the only other exceptions. Several measures achieved values that were ≥.9, including the Intolerance of Uncertainty Scale (IUS), Obsessive-Compulsive Inventory-Revised (OCI-R), PTSD Symptom Checklist (PCL), Social Phobia Scale (SPS), and White Bear Suppression Inventory (WBSI).

In order to examine differences between the scores obtained by participants with and without a given diagnosis effect sizes were calculated for each measure based on the formula outlined by Hasselbad & Hedges (1995), which compares the standardized distance between mean scores obtained by each diagnostic group (i.e., participants with and without a positive classification for each of the diagnostic categories predicted in the current study; See Table 4). Recent methodological literature has consistently described the effect size statistic as being

superior to *p*-values for determining differences when utilizing ROC analyses using the signal detection framework (for detailed review see Kraemer & Kupfer, 2006).

3. Diagnostic Efficiency

Independent sets of analyses were conducted for each diagnostic classification predicting participants' diagnostic status using all self-report scores. For the first set of ROC analyses, participants were grouped into those who met diagnostic criteria for at least one psychological disorder and those who did not; the second set grouped participants on the basis of having ≥ 2 diagnoses; and the remaining eight sets of analyses predicted specific diagnoses individually (outlined previously in METHODS section).

Outcome of diagnostic utility examinations are summarized in Tables 6 through 15. Each table provides multiple indices of measures' performance when used to predict participants' diagnostic status within each category (outlined in Table 5). The current study erred on the side of comprehensiveness concerning which statistical indices to report; however, it is important to note that differential importance is placed on the statistical indices that are cited as superior in the most up-to-date methodological literature in this domain (i.e., kappa, negative/positive predictive values, etc.; Kraemer, 2013; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Zhou, Obuchowski, & McClish, 2002) over those that have been highlighted as possibly misleading or lacking clinical utility (Nickerson, 2000; Streiner & Norman, 2006; Gigerenzer et al., 2007; Wegwarth, Schwartz, Woloshin, Gaissmaier, & Gigerenzer, 2012; Krämer & Gigerenzer, 2005; Kraemer, 2013). Due to the comprehensiveness of the results only the outcomes most pertinent to the research questions addressed in the current study are reported and/or interpreted within the text. Tables 6 through 15 provide numerical values for all outcomes

produced in current study and these values that can be used to answer independent research questions regarding psychometric performance and/or diagnostic efficiency of the measures. Additionally, the author can be contacted for further clarification if needed.

Following is a broad overview of the performance of the optimal measure for each diagnostic category as identified by ROC analyses. Measure performance within diagnostic category was examined and subsequent determination of the best performing measure was based on the weighted Cohen's kappa statistic (*w*= .5; Cohen, 1968) and the Q-ROC curve (Kraemer, 1992). The best performing measures for each diagnostic category and relevant statistical indices are presented in Table 15. Results are interpreted using the terminology and presentation format that has empirically demonstrated ability to facilitate interpretability of medical statistics (i.e., Gigerenzer & Edwards, 2003; Krämer & Gigerenzer, 2005; Gigerenzer et al., 2007)

≥1 Disorder (Table 6). When detecting the presence of any psychopathology (i.e., ≥1 disorder; Prevalence = .661) the Anxiety Sensitivity Index (ASI-3) emerged as the best measure overall with an AUROC value of .687. At the optimal cut-point of seven, 68.8% of the total sample was accurately classified (i.e., *efficiency*). In terms of sensitivity and specificity, 76.7% of cases classified as positive were *true* positives and 53.5% of cases classified as negative were *true* negatives. Based on the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) statistics, 76 out of 100 individuals who obtain a score ≥ 7 will have at least one psychiatric disorder (of those assessed in the current study) while 54 out of 100 individuals who fail to reach this cut-off (i.e., score of 7) will not have any psychiatric disorders present. Therefore, the ASI-3 was better at predicting who did have at least one psychiatric disorder (i.e., true positive cases) present versus identifying those who did not (i.e., true negative cases).

≥2 Disorders (Table 7). A cut-point of 11 on the DASS Total (sum of the three subscales) most efficiently predicted the presence of two or more psychological disorders (i.e., *≥2 Disorders;* Prevalence= 37.7%) accurately classifying 69.5% of the sample. Overall, the DASS Total index test produced an AUROC value of .674. Of those classified as 'positive' 63.6% were 'true positives' and of those classified as 'negative' 73.1% were 'true negatives.' Based on the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) statistics, 59 out of 100 individuals who obtain a score ≥ 11 will have two or more psychiatric disorders (of those assessed in the current study) while 77 out of 100 individuals who fail to reach this cut-off (i.e., score of 11) will not have two or more disorders present. Therefore, the DASS was better at predicting who did *not* have two or more psychiatric disorders (i.e., true negative cases) present versus identifying those who did (i.e., true positive cases).

Major Depressive Disorder (MDD; Table 8). The Depression Subscale of the DASS emerged as the best predictor of a Major Depressive Disorder (MDD) diagnosis (Prevalence= 6.5). A cut-point of 11 produced an AUROC estimate of .760 and accurately classified 91.4% of the sample. The sensitivity at this cut-point was much lower than the specificity (.263 versus .960, respectively), meaning the Depression Subscale of the DASS was better at screening out those who do *not* have depression than identifying those who do possess this diagnosis. Accordingly, 31 out of 100 individuals who score an 11 or higher on the DASS Depression Subscale will have a positive diagnosis of MDD (i.e., PPV) and, of those whose score falls below 11, 95 out of 100 will not (i.e., NPV). Comparison of the DASS Total score and the DASS Depression subscale revealed no significant differences ($z = .10$, $p \leq .919$) in ability to predict a diagnosis of Major Depressive Disorder (using Hanley and McNeil's [1983] method for comparing dependent ROC curves).

Generalized Anxiety Disorder (GAD; Table 9). The DASS Total emerged as the best predictor of a GAD diagnosis (Prevalence= 18.5%) achieving an overall AUROC value of .800. The optimal cut-point was a score of 11, which accurately classified 72.3% of the sample. Obtained value for sensitivity was higher than specificity (.852 versus .693, respectively) meaning that a cut-point of 11 on the DASS was more effective at identifying positive cases than negative and resulted in overidentification (i.e., high frequency of false positives). Accordingly, only 39 out 100 individuals with a score of 11 on the DASS will have a diagnosis of GAD (i.e., PPV), while 95 out of 100 score below this cut-point will not (i.e., NPV).

Obsessive Compulsive Disorder (OCD). Results of the ROC analyses predicting diagnosis of OCD (Prevalence = 1.7%) did not produce any significant results due to inadequate sample size (i.e., few identified cases meeting diagnostic criteria for OCD)[1]. Methodological guidelines dictate that at least ten participants in a given sample must have a true positive diagnosis (as labeled by the gold standard assessment method) in order to apply ROC methodology appropriately (Kraemer, 1992). Any results derived from calculations on these data would thus be misleading, and were thus not reported in the text or tables.

Panic Disorder (PD; Table 10). The Intolerance of Uncertainty Scale (IUS) emerged as the best predictor of a Panic Disorder (PD) diagnosis (Prevalence= 7.2%). Overall, the IUS produced an AUROC estimate of .695 and a cut-point of 63 accurately classified 93.5% of the sample. The sensitivity at this cut-point was lower than the specificity (.286 versus .985, respectively) meaning the IUS was better at screening out those who do *not* have difficulties with panic than identifying those who do possess this diagnosis. Accordingly, 60 out of 100

[1] The statistical software (i.e., ROC-5 program; http://mirecc.stanford.edu) used to conduct analyses in the current study will not predict classification when sample size of positive cases falls below the minimum threshold ($n < 10$ true positive cases).

individuals who score a 63 or higher on the IUS will have a positive diagnosis of PD (i.e., PPV) and, of those whose score falls below 63, 95 out of 100 will not (i.e., NPV).

Post-Traumatic Stress Disorder (PTSD; Table 11). The PTSD Symptom Checklist (PCL) score emerged as the best predictor of a PTSD diagnosis (Prevalence= 4.5%) with an AUROC value of .873. The optimal cut-point was a PCL score of 31, which accurately classified 92.8% of the sample. The sensitivity at this cut-point was lower than the specificity (.500 versus .950, respectively) meaning the PCL was better at screening out those who do *not* have PTSD than identifying those who do possess this diagnosis. Accordingly, only 33 out of 100 individuals with a score of 31 on the PCL will have a diagnosis of PTSD (i.e., PPV), while 97 out of 100 score below this cut-point will not (i.e., NPV).

Social Anxiety Disorder (SAD; Table 12). The Behavioral Inhibition Scale (BIS) emerged as the best predictor of a diagnosis of Social Anxiety Disorder (Prevalence = 40.7%) with an AUROC value of .656, accurately classifying 65.8% of the sample at the optimal cut-point of 15. The sensitivity at this cut-point was lower than the specificity (.580 versus .711, respectively) meaning the BIS was better at screening out those who do *not* have Social Anxiety Disorder than identifying those who do possess this diagnosis. Accordingly, only 58 out 100 individuals with a score of 31 on the BIS will have a diagnosis of Social Anxiety Disorder (i.e., PPV), while 71 out of 100 score below this cut-point will not (i.e., NPV).

Specific Phobia (SP; Table 13). The Anxiety Sensitivity Index-3 (ASI-3) emerged as the best predictor of a Specific Phobia diagnosis (Prevalence= 41.4%) producing an AUROC value of .623 and accurately classifying 61.6% of the sample at a cut-point of 10. Obtained value for sensitivity was slightly higher than specificity (.636 versus .602, respectively) meaning that a cut-point of 10 on the ASI-3 was more effective at identifying positive cases than negative and

resulted in a slight rate of overidentification (i.e., false positives). Fifty-three out of 100 individuals with a score of 10 on the ASI-3 will have a diagnosis of Specific Phobia (i.e., PPV), while 70 out of 100 score below this cut-point will not (i.e., NPV).

Eating Disorders (ED; Table 14). The Eating Attitudes Test-26 (EAT-26) emerged as the best predictor of Eating Disorder diagnostic status (Prevalence= 12.0%) with an AUROC value of .569, overall, accurately classifying 87.7% of the sample using a cut-point of 27. The sensitivity at this cut-point was lower than the specificity (.143 versus .977, respectively) meaning the EAT-26 was far better at screening out those who do *not* have an Eating Disorder than identifying those who do possess this diagnosis. Accordingly, only 45 out of 100 individuals with a score of 27 on the EAT-26 will have a positive Eating Disorder diagnosis (i.e., PPV), while 89 out of 100 score below this cut-point will not (i.e., NPV). These results should be interpreted with caution, however, given homogenization of different diagnoses (i.e., anorexia nervosa, bulimia, and eating disorder not otherwise specified), a low base-rate, and issues with reliability of the measure in general (Garner et al., 1982).

Practical implications of obtained results are outlined in DISCUSSION to follow (i.e., Gigerenzer & Edwards, 2003; Krämer & Gigerenzer, 2005; Gigerenzer et al., 2007).

IV. DISCUSSION

1. Conclusions

This study sought to identify the diagnostic accuracy of brief, self-report instruments in a non-clinical sample using diagnoses based on structured interviews as referents. Following Kraemer's (1992) guidelines for evaluating medical tests, multiple methods were used to evaluate diagnostic accuracy and results were reported in line with the recommendations outlined in the STARD criteria (Bossuyt, et al., 2003).

Index test performance. As expected, ROC analyses revealed significant overlap in terms of measures' predictive validity across diagnostic categories. Specifically, the Depression Anxiety and Stress Scales (i.e., DASS) emerged as the best measure in three of the nine sets of analyses examining diagnostic utility (i.e., ≥ 2 Disorders, Generalized Anxiety Disorder, and Major Depressive Disorder). The Anxiety Sensitivity Inventory-3 (ASI-3) was the next best performing measure emerging as the most efficient predictor for two of the nine diagnostic categories (i.e., ≥ 1 Disorder and Specific Phobia). The two diagnostic categories best predicted by the ASI-3 had a high degree of overlap given the high prevalence of Specific Phobia diagnoses in the current sample and, specifically, 62.7% of participants who fell into the ≥ 1 Disorder category had a diagnosis of Specific Phobia. Therefore, the generalizability of the ASI-3 that could be interpreted from overview results may be based on its prediction of largely the same criterion across two different sets of analyses rather than broad clinical applicability (Fan, Thompson, & Wang, 1999).

Overall, there were also measures that demonstrated consistently poor performance across most diagnostic categories used for prediction. One such measure was the Discomfort Intolerance Scale (DIS), which demonstrated poor predictive validity for all diagnostic categories except Panic Disorder. The DIS also produced the second lowest value for internal consistency observed in the current study with a Cronbach's alpha of .466 (with the lowest being the EAT-26; Cronbach's alpha = .287). Analyses using the DIS for prediction produced low values for global measures of diagnostic utility (i.e., AUROC, kappa, efficiency, etc.) and did not perform better than prediction based on chance alone (i.e., non-significant chi-squared values produced). The DIS demonstrated somewhat better diagnostic utility (i.e., AUROC = .677; significantly better than chance at the $p \leq .001$ level) when predicting the presence of Panic Disorder, however, which is consistent with the theoretical underpinnings of the measure intended to assess subjective discomfort in response to physiological sensations associated with negative affect (Schmidt, Richey, & Fitzpatrick, 2006). The finding that the DIS performed optimally when predicting Panic Disorder may speak to the discriminant validity of this measure and point to the need for measure refinement efforts in order to capitalize on this strength.

Predicting specific diagnoses. Ability of measures to predict diagnostic status for specific disorders was widely discrepant overall. Most index tests designed to measure symptoms associated with a specific disorder demonstrated inferior performance (relative to other non-specialized index tests) when predicting their respective diagnostic target. In fact, only three of the measures specifically designed to measure domain-specific symptoms of the outcome criterion (i.e., psychiatric diagnosis) emerged as the best predictors of their respective diagnoses: the PTSD Symptom Checklist (PCL) predicting PTSD; the Depression Subscale of the DASS predicting Major Depressive Disorder; and the Eating Attitudes Test-26 (EAT-26) detecting the

presence of Eating Disorders. Although there were issues with reliability of the EAT-26, evidence for the utility of the other two instruments was also supported by discriminant examinations. Specifically, the PCL and the Depression subscale of the DASS demonstrated the highest levels of discriminant validity overall (i.e., performing poorly when used to predict other diagnostic classifications for specific disorders) and also achieved good to excellent values for internal consistency (.941 and .855, respectively; Cronbach, 1951). While the EAT-26 outperformed other index tests in terms of efficient prediction of Eating Disorders (accurately classifying 87% of cases) it demonstrated a high rate of false positives with only 14.3% of cases identified as positives actually meeting diagnostic criteria according to the gold standard. However, the low rate of false negatives (i.e., high specificity value) achieved by the EAT-26 may point to this measure's ability to function effectively as a device for "screening out" non-eating disordered individuals (97.7% of cases labeled negative were true negatives).

These results are consistent with what would be expected given existing recommendations for appropriate application of the EAT-26 (Dotti & Lazzari, 1998; Siervo, Boschi, Papa, Bellini, & Falconi, 2005). The original development article advised against using the EAT-26 as a diagnostic tool, instead promoting its appropriateness as a screening device (Garner et al., 1982; Berland, Thompson, & Linton, 1986; Beidas et al., 2014). Research has generally supported the EAT-26's utility as a screening device and appropriateness for application in practice settings (Anderson & Paulosky, 2004). The pattern of results achieved by the EAT-26 suggests that non-disordered individuals may screen positively and be classified inaccurately, but disordered individuals are not likely to be falsely labeled and thus overlooked (Zimmerman, 2009; Streiner & Cairney, 2007; Streiner & Norman, 2003; 2006). Therefore, while not optimal in terms of discrete categorization, the diagnostic utility outcome of the EAT-

26 for prediction of Eating Disorders is acceptable for application as a screening device given that the costs associated with failing to identify a positive case (i.e., decreased likelihood that a disordered individual will receive any treatment or be accurately diagnosed in the future) are far more consequential than incorrectly labeling a negative case as positive (i.e., unnecessary application of further evaluation procedures).

Regarding outcome for the other diagnosis-specific measures used in the current study, the overall results suggested that instruments generally demonstrated better sensitivity than specificity, resulting in many false positive cases. This pattern of results makes sense if the vast majority of individuals with a positive diagnosis endorsed scale items and individuals without a positive diagnosis also endorsed items to some degree. This could be a function of measuring non-pathological features of a psychological disorder that are not included in the "gold-standard" diagnostic criterion on which diagnostic label is based (Blashfield, 1990; Clark et al., 1995; Andrews, 1996; Bogenschutz & Nurnberg, 2000). Other influential factors may include low levels of reliability (i.e., Cronbach's alpha) observed within scales (Table 5).

Several of the measures designed to measure symptoms of a specific diagnosis performed well on certain outcome domains but failed to perform *efficiently*. For example, a cut-point of 23 on the Anxiety Sensitivity Inventory-3 (ASI-3) produced a higher AUROC value than the measure identified as the optimal predictor of a Panic Disorder diagnosis (i.e., Intolerance of Uncertainty Scale [IUS]; .712 versus .695, respectively). Such discrepancies in performance were evident across several diagnostic categories and are, in general, attributable to differential performance balancing accuracy for identifying both positive cases *and* negative cases (i.e., higher efficiency value). In the aforementioned example the ASI-3 correctly identified only

82.5% of the population, while the Intolerance of Uncertainty Scale (IUS) correctly labeled 93.5% of the sample.

Taken together, these findings support the use of several of several self-report measures as screening devices under *some* circumstances (Streiner & Norman, 2006). Some measures performed well for identifying *non-disordered* cases (i.e., high NPV), others were better at identifying *positive* cases (i.e., high PPV), and some performed *efficiently* (i.e., high value for efficiency) accurately classifying a large proportion of negative and positive cases within the sample as a whole (i.e., optimal balance of NPV and PPV). The costs associated with inaccurate classification must be weighed in order to determine if incremental information provided by a screening device will have a positive impact, which is determined by the goals of a particular setting (Griffith, 1997; Brown & Davis, 2006). For example, if a screening device is applied in isolation for determining who should receive psychotherapeutic intervention then the costs associated with an individual not receiving treatment who needs it must be weighed against incurred by proving treatment to someone who does not need it (i.e., false positive; Lang & Stein, 2005). Research has supported the conclusion that no harm is induced by treating (i.e., using an evidence-based, diagnosis-specific psychological intervention) an individual whose symptoms are, in actuality, sub-threshold (i.e., does not meet diagnostic criteria; Borkovec & Ruscio, 2001). Therefore, in this scenario, selection of a measure that performs well when screening out non-disordered individuals might be most appropriate (Deeks, 2001; Behar, Alcaine, Zuellig, & Borkovec, 2003).

It is important to note that not all outcome-measure research is created equal, particularly in terms of adhering to reporting standards and promoting external validity of results (Kessel & Zimmerman, 1993; Haynes, Nelson, & Blaine, 1999; Bossuyt, et al., 2003; Kraemer, 1992;

Cook, Cleland, & Huijbregts, 2007; Reid, Lachs, & Feinstein, 1995; Whiting et al., 2004). A common complaint concerning diagnostic outcome research is the failure of researchers to report incremental information concerning measure performance beyond values for AUROC, sensitivity, and specificity (Streiner & Norman, 2006; De Vet et al., 2001; Kraemer, 2013). Additionally, the determination of which measure is "best" is generally made solely based on the AUROC value, which can lead to inaccurate, biased conclusions and conceal important nuances of measure performance (Wilczynski, 2008; Selman, Morris, Zamora, & Khan, 2011). For example, while some measures in the current study did not perform optimally on indices of overall performance (e.g., low AUROC value) they often demonstrated superior performance relative to most efficient measures in other performance domains (e.g., high NPV). If outcome research adheres to published methodological standards and recommendations for reporting of medical statistics (i.e., Gigerenzer et al., 2007; 2010; Wegworth et al., 2012), the incremental information that results can facilitate interpretability allowing individualized selection of the most appropriate screening index and relative cut-off guided by desired outcome, which may not be "overall" performance (Eisman et al., 2000; Garb, 2003; Wilczynski, 2008).

The finding that measures specifically developed for assessment of symptoms that characterize a given diagnosis performed relatively poorly as compared to other instruments included in the current study's battery cannot be solely attributed to poor diagnostic validity (Hudziak et al., 2004). Other factors besides diagnostic concordance likely influenced diagnostic validity outcomes reported in the current study including prevalence rates, measure performance (e.g., internal consistency), less sophisticated interviewers than those utilized in epidemiological studies (i.e., Kessler et al., 2005), and characteristics of the analogue sample examined (i.e., likely accounting for observed floor effects). However, the degree of influence exerted by each

of the aforementioned variables cannot be quantified using the outcome data produced in the current study. Therefore, the results of the current study, taken alone without a more nuanced application of theory, replication of results, and balance of all performance factors, would lead to an inaccurate conclusion. Independent replication of the current study in both similar and divergent samples is necessary to answer research questions of this nature (Zweig & Campbell, 1993). Evaluation of the other (harder; more esoteric) output is also necessary in line with contemporary methodological standards (i.e., STARD criteria; Bossuyt, et al., 2003; Kraemer, 1992) in order to arrive at a practical, empirical conclusion likely to be useful in applied settings (Basco et al., 2000; Grubaugh et al., 2007).

Results were consistent with the Tripartite Theory framework (i.e., Clark & Watson, 1991) and the extensive body of empirical research demonstrating significant overlap of depression and anxiety symptoms, thus providing further evidence that various psychiatric diagnoses are differential manifestations of the same underlying pathology (Andrews, 1996; Frances et al., 1990; Brown et al., 1997; 1998). Treatment research applying a transdiagnostic approach that targets the common features of emotional disorders (i.e., *Unified Protocol;* Allen et al., 2008; Barlow et al., 2010; Ellard, Fairholme, Boisseau, Fairholme, & Barlow, 2010; Fairchione et al., 2012) has produced equal efficacy results to disorder-specific psychosocial interventions with established efficacy (i.e., Chambless & Hollon, 1996; 1998). Interestingly, the measure that demonstrated the highest degree of overlap in terms of cross-diagnostic predictive ability and best overall performance in the current study (i.e., the Depression Anxiety Stress Scale; Brown et al., 1997) was developed in accordance with the theoretical underpinnings of the Tripartite Theory framework (Clark & Watson, 1991). The results of the current study are in line with accrued research supporting the internal and external validity of the DASS since its

inception (i.e., Antony et al., 1998; Clara, Cox, & Enns, 2001; Henry & Crawford, 2005; Ng et al., 2007; Norton, 2007; Crawford et al., 2009; Osman et al., 2012) demonstrating superior ability to predict presence of ≥ 2 Disorders, Generalized Anxiety Disorder, and Major Depressive Disorder to that demonstrated by other measures used for prediction.

2. Practical Implications

In an effort to streamline, reduce costs, and enhance comprehensiveness of treatment psychiatric services have recently expanded to non-traditional settings not specialized for mental health provision (Hatfield & Ogles, 2007). Accordingly, the spectrum of healthcare providers involved in the diagnosis and treatment of psychological disorders has broadened as well (Lang & Stein, 2005). However, given the difficulty that specifically-trained *doctoral*-level mental healthcare specialists display arriving at the same diagnostic conclusions (in the absence of a systematic assessment procedures; Oskamp, 1962; Grove et al., 2000; Garb, 2005), paramount importance should be placed on the routine application of standardized, data-driven diagnostic procedures in non-traditional settings – assessment in many of which will often be implemented by paraprofessionals (Faust et al., 1988; Turner, 1966; Garb, 1989; Clark, et al. 1995; Basco et al., 2000; Brammer, 2002). Information derived from self-reports designed to assess symptoms targeting specific disorders can aid in this task decreasing the amount of clinical inference (i.e., clinical intuition) required to arrive at diagnostic conclusions (Garb, 1994; Grubaugh et al., 2007). The current pattern of results, however, suggests that the set of widely disseminated, freely available, strongly psychometrically supported instruments examined are not optimal for this task, and thus additional future research to improve predictive values is warranted (i.e., Garb, 2003). Nonetheless, the application of this group of measurements affords considerable

advantage over a diluted version of clinical judgment in contexts where few or no mental healthcare professionals work (Gadol, 1969; Beidas et al., 2014).

Measures used for prediction in the current study demonstrated variable performance across diagnostic validity outcomes. For example, several measures accurately identified the majority of the negative cases (i.e., high *negative predictive power*) but performed poorly when identifying the positive cases (i.e., low *positive predictive power*). These quantitative outcomes translate to differential practical utility for answering specific clinical questions in practice (Streiner & Norman, 2006; Kraemer, 2013). High negative predictive power speaks to measures' ability to accurately 'screen out' individuals (i.e., low rate of false negatives) and would be clinically useful when making determinations concerning who should be more thoroughly evaluated (i.e., belying confidence that the patient is not experiencing symptoms of the corresponding disorder; Coyne & Schenk, 1997; Basco et al., 2000). Therefore, such measures would be appropriate for use in a setting with abundant resources that can afford to deliver in-depth evaluation procedures to a larger group of individuals in an effort to avoid missing positive cases (i.e., assigning differential importance to *specificity*; Schulberg et al., 1985). In the context of psychiatric screening failing to identify a positive case carries more severe and far-reaching consequences than incorrectly labeling a positive case as negative (Kraemer, 2003; Zimmerman, 2010). The cost of subjecting all patients (or even *most*) to the "gold-standard" evaluation procedure is prohibitive (i.e., time intensive; too few skilled interviewers; expensive; etc.) making a screening approach an attractive, feasible alternative when attempting to strike the optimal "cost-benefit" ratio (Bossuyt et al., 2003; Grubaugh et al., 2007).

To expand on this point when considering multiple healthcare delivery settings, the cost-benefit ratio of expanding services is determined by the quality of services provided (Yates &

Taub, 2003). Thus the increased availability and access to psychiatric services is only desirable if incremental improvement in patient outcome results (Basco et al., 2000; Garb, 2003). Desirable patient outcome (i.e., reduced symptoms; greater function; convergence of actuarial and clinical indexes of change) is determined overall by whether or not a given individual receives the correct treatment, the selection of which is most often guided by initial diagnosis (Mellor-Clark et al., 1999). Different disorders, even those in the same class, confer differential treatment recommendations; therefore, the process of choosing the correct treatment strategy is highly dependent upon accurate classification prior to treatment planning (Lambert et al., 2003; Zimmerman et al., 2010). This issue of differential diagnosis is an especially difficult task in the context of medical-oriented settings (e.g., primary care) due to the similar presentation of anxiety symptoms and various medical disorders (e.g., physiological arousal in panic disorder; Coyne et al., 1995; Löwe et al., 2008). Even if patients' symptom presentation is correctly attributed to an anxiety-related disorder the medical provider (i.e., one who is typically not specifically trained in psychiatric taxonomy and/or familiar with evidence-based treatments) must then decide which specific anxiety disorder is warranted (Lang & Stein, 2005; Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007). If a patient is diagnosed and treated for a disorder with which he or she is not actually afflicted, not only are time and resources wasted, but iatrogenic effects of incorrect treatment application may also occur (Edwards, 2003; Hatfield & Ogles, 2004; Mitchell et al., 2009). The methodological approach utilized in the current study allows differential importance to be applied to sensitivity or specificity (Kraemer, 1992)- a decision based on setting-specific characteristics including the population served (e.g., high base-rates of a particular disorder), systemic goals (e.g., limiting systemic liability), and/or availability of resources (e.g., trained clinicians; Meyer et al., 1998). Understanding the properties and utility of

simple, freely available, scientifically-sound instruments such as those examined in this study may reduce the degree of intuitive clinical decision-making in disparate settings, and ultimately be of potential benefit to modern efforts to more closely integrated healthcare across scopes of practice (Garb, 1995; Haynes & Lench, 2003; Kraemer, 2003; Lambert et al., 2003; Beidas et al., 2014).

The current study also yielded insight into the complexity of conveying information to clinical service providers, particularly those with limited training and practice focusing on mental health issues (Garb, 1994; Lambert et al., 1998). Given the pragmatic nature of the ultimate goals of the current study (i.e., diagnostic screening in practice settings) some discussion of this issue is relevant. For example, the overall pattern of results, comprising numerous (dense) tables to arrive at general conclusions is not easily communicated to a clinical provider with limited training and/or interest in actuarial prediction (Gigerenzer, 2002; Gigerenzer et al., 2007; Edwards, 2003). It should be pointed out that this rather onerous amount of information (in reference to the applied domain) exists in the context of a conscious decision to use the most parsimonious, focal factor structure supported for research in diagnostic prediction (Gigerenzer, Wegwarth, & Feufel, 2010). A more nuanced view of the instruments examined would likely have afforded greater parsing of domains and configuration of predictor vs. criterion variables; however, these methods would have been antithetical to the goal of dissemination to practice settings (Lambert et al., 1998; Glasgow et al., 2003). Decades of interdisciplinary research demonstrate that the complexity of research products is negatively associated with likelihood of adoption and implementation (see Rogers, 2004; 2010). By way of tangible example, the question in practice is unlikely to be a variation on the often-repeated academic theme of “yes, but have you also thought about…?” and more likely to be focused on

"which one of these measures works best to help the patient I'm seeing today?" (Gigerenzer & Edwards, 2003; Zimmerman et al., 2010).

3. Limitations and Future Directions

The current study erred on the side of comprehensiveness concerning which statistical indices to report. For example, to be consistent with available measure-validation literature the "p value" statistic was reported for comparisons across measures; however, as Kraemer (2003) points out these computations have no real utility for measurement validation purposes (for detailed explanation of "p value" misuse in research see: Wilkinson, 1999; Dar, Serlin, & Omer, 1994), particularly in contrast to measures of discriminant effect sizes between clinical vs. non-clinical groups (also reported). Another potential limitation of this study is the usage of a non-clinical sample whose diagnostic base-rates closely resemble those of the general population. As Kraemer (1992) and Smith (1995, pg. 37) point out, indices of diagnostic validity can change when the same index test is used in different settings, and the environment examined is somewhat insular with regard to likely generalizability to a broader population (i.e., potentially limited external validity in extrapolation of these conclusions; Kraemer, Kupfer, Clarke, Narrow. & Regier, 2012). Therefore the suggested cut-offs presented in this study would be appropriate for individuals in similar settings (e.g., college campus in the cultural context of the Southern United States), but may not be appropriate in clinical contexts reflecting inflated base rates of psychopathology (e.g., inpatient settings). Alternatively, this methodology and accessing an analogue (i.e., "normal") population have also been cited as significant design strengths in burgeoning research examining differential prediction of clinical instruments (i.e., paired design approach; Kraemer, 1992). Regardless, these results are unlikely to generalize across settings,

and future research examining the diagnostic utility of these instruments in other environments, particularly those providing clinical services, is warranted (Streiner & Norman, 2006; Kraemer, 2013).

Additionally, low base-rates observed for Obsessive-Compulsive Disorder (OCD) and specific forms of Eating Disorders (i.e., Anorexia, Bulimia, and Eating Disorder Not Otherwise Specified; See footnote in Table 1 for respective sample sizes) precluded application of ROC analytic methodology as applied for prediction of other diagnoses in the current study. The statistical software (i.e., ROC-5 program; http://mirecc.stanford.edu) used to conduct analyses in the current study is unable to predict classification when sample size of positive cases falls below the minimum threshold ($n < 10$ true positive cases), which is consistent with sample size requirements outlined in methodological literature (i.e., Kraemer, 1992) and non-parametric methods more generally (i.e., Tabachnick & Fidell, 2012). This sample size requirement precluded analyses examining the diagnostic utility of index tests for predicting OCD diagnoses since only five participants were assigned this diagnosis based on the gold standard (i.e., SCID-II diagnostic interview).

Despite insufficient sample size for individual ROC analyses, positive OCD diagnoses were included in analyses predicting *any* positive diagnostic classification (i.e., ≥ 1 and ≥ 2 diagnoses of psychopathology) since observed base rates did not deviate substantially from those reported for community samples (Kessler et al., 2005). Additionally, all descriptive statistical outcomes regarding OCD diagnostic status are reported in the current study. This approach models that applied in similar diagnostic accuracy research (i.e., Youngstrom et al., 2004) and was intended to optimize the amount of information extracted from the current sample's data while still adhering to methodological guidelines (Kraemer, 1992; 2013).

Likewise, examinations of particular forms of eating disorders were also limited on the basis of their occurrence. Due to the low prevalence of anorexia nervosa (n=4), for example, all eating disorder diagnoses assigned in the current sample (which included bulimia nervosa [n=14] and eating disorder not otherwise specified [n=17]) were collapsed into a single "eating disorder" category (n=35) in order to achieve adequate sample size for applying ROC analyses (without excluding diagnosis of anorexia nervosa). In addition to conforming to methodological approaches utilized in similar studies predicting low base rate diagnoses (i.e., Youngstrom et al., 2004; Basco et al., 2000) this strategy also comports with the eating-disorder symptoms self-report measure (i.e., EAT-26) used for prediction in the current study, which focuses on general symptoms as opposed to diagnosis-specific behaviors (Dotti & Lazzari, 1998; Siervo et al., 2005).

The results of these analyses examining aggregated eating disordered outcomes indicated that few instruments offered great utility in prediction of diagnosis, with cut points across measures often set to extreme scores (i.e., 0 or no endorsement of symptoms). Although discouraging for clinical application, this pattern of results potentially informs the need for future research developing new measures with a focus on constructs differentially related to eating pathology specifically (e.g., overly attentive to weight and shape; preoccupation with food; ritualization of eating behaviors; caloric restriction; bingeing or purging; etc.; Hudson, Hiripi, Pope, & Kessler, 2007) as opposed to more general processes relevant to a broader array of pathologies. Similarly, these future efforts could focus on development of domain and symptom-specific measures with greater internal reliability, as the EAT-26 (one of the most widely distributed self-reports for ED symptoms; Beidas et al., 2014) has notably demonstrated poor reliability of some of its subscales since its development (Garner et al., 1982; Berland et al.,

1986). This consistently-noted issue was problematic in the current study as well, which likely limited conclusions that could be formed concerning prediction of eating disorders.

Additionally, the rates of positive cases of pathology in the current sample may have led to discrepant conclusions, particularly when considering external validity of these results (Kraemer, 2003; Kraemer et al., 2012). The majority of the sample met diagnostic criteria for at least one disorder (66.1%), slightly more than one third met criteria for two or more disorders (37.7%), and one third (33.6%) did not meet criteria for any disorders assessed. Taken together, these overall base-rates are inflated as compared to available published epidemiological reports of the prevalence of psychopathology among the general population (i.e., Kessler et al., 2005), which (as previously pointed out) affects the outcome of ROC analysis (i.e., ideal cut-point, performance indices, etc.; Kraemer, 1992). Observed rates of Specific Phobia (SP) were particularly discrepant from base rates estimated in the American population at large, with fully 41.3% of the current sample meeting diagnostic criteria for SP. In comparison, the National Comorbidity Replication Study (Kessler et al., 2005) reported an estimated point prevalence rate of 13.3% in a sample similar in age to that of the current study. Observed rates of Social Anxiety Disorder (i.e., SAD) in the current sample were also inflated as compared to those observed in the National Comorbidity Replication sample (40.7% versus 13.6%, respectively; Kessler et al., 2005; Ruscio, Brown, Chiu, Sareen, Stein, & Kessler, 2008).

These elevated base-rates could have been attributable to the current study's utilization of relatively less trained raters in comparison to many actuarial examinations of clinical instrumentation (i.e., undergraduate research assistants with more limited training and little to no clinical experience) than those utilized in Kessler's studies (i.e., paid, experienced interviewers whose interview fidelity was subjected to a greater degree of scrutiny; Kessler et al., 2004;

2008). Additionally, over-reporting on the part of research participants may have played a role (Antony et al., 1994). Specifically, responses to questions pertaining to specific phobia and social anxiety disorder, particularly aversion to environmental stimuli and/or situations, were likely subject to some interpretation on the part of the interviewer (Garb, 1994; 1995; 1997; 2005; Kendler, Karkowski, & Prescott, 1999). These are also symptoms of a high base rate without implications for functional impairment or diagnosable psychopathology (Regier et al., 2008; Woody & Teachman, 2000; Bögels et al., 2010). For example, endorsement of a high degree of fear of public speaking is not only common, but *normative* across the lifespan (e.g., Heimberg, Hope, Dodge, & Becker, 1990; Stein, Walker, & Ford, 1994), but is associated with additional symptoms indicative of diagnosis in only a small subgroup of people afflicted (Ruscio et al., 2008; Blöte, Kint, Miers, & Westenberg, 2009). Further, research has demonstrated tendency of individuals to conflate aversive emotional responses to environmental stimuli, particularly when delineation between *disgust* versus *fear* is necessary (i.e., Muris, Mayer, Huijding, & Konings, 2008). In the event that respondents rated some of their reactions to situations and/or stimuli as severe (i.e., pathological), the relatively less-trained raters were left with no choice but to record their answers verbatim (Oskamp, 1962; Garb, 2005). A more experienced or specifically trained clinical interviewer may not have asked additional questions regarding functional impairment and offered a more accurate depiction of diagnosis (Martin, Horder, & Jones, 1992; Muris, Merckelbach, de Jong, & Ollendick, 2002). Differential methods of assessment utilized in the current study versus the Comorbidity Replication Study (Kessler et al., 2005) may have also contributed somewhat to the observed diagnostic disparity (Garb, 2003). Notably, Kessler and colleagues' (2005) study used a different structured diagnostic interview as their diagnostic criterion (i.e., World Mental Health Survey Initiative Version of the

World Health Organization Composite International Diagnostic Interview [WMH-CIDI]; Kessler & Üstün, 2004) than those utilized in the current study (i.e., the ADIS-IV and the SCID); however, these two diagnostic interviews use the same reference criteria (i.e., DSM-IV) for diagnosis, suggesting that observed differences are not likely not an artifact of criterion-related bias (Welner, Liss, & Robins, 1974; Basco et al., 2000), and thus any discrepancies would be a function of probability of error based on internal validity estimations for each instrument (Kraemer, 2013).

Albeit inflated, the observed base-rates of psychopathology in the current sample suggest that mental health is indeed a concern on college campuses and support the expansion and increased presence of mental health services on college campuses (Kadison & DiGeronimo, 2004; Mowbray et al., 2006). Too often mental health problems are only identified in hindsight after the occurrence of an adverse event (e.g., suicide attempt, violent behavior, drop in school performance, friends becoming concerned about uncharacteristic social isolation, etc.) and unfortunately it may be too late to intervene at this point (Garlow et al., 2008; Cook, 2007). A feasible approach to proactive identification might be to require quick mental health screenings at various time points throughout the year, perhaps at time points that have been identified as being especially hard (i.e., the initial transition to college; final exams; graduation; etc.; Voelker, 2003).

Accrued evidence from interdisciplinary research on variables that affect the diffusion process (e.g., Rogers, 2010) can be used to enhance success of such implementation efforts. For example, identifying and providing solutions to system-specific problems and/or concerns that fit with the system's organizational structure has often been shown to enhance the acceptability and perceived utility of a proposed innovation (Zimmerman et al., 2010). This is especially true when

stakeholders at multiple levels are involved and/or affected (Rogers, 2004). Successful implementation could be facilitated by integrating mental health screening into a preexisting data gathering process (e.g., teacher evaluations conducted at the end of the semester or surveys conducted during mandatory activities for attendance, such as health clearance appointments; Kitzrow, 2003). This would not only streamline data collection and reduce potential barriers to accessibility, but could also allow provision of immediate feedback to students (Cook, 2007; Mowbray et al., 2006). This feedback could be useful to guide treatment referral for identified difficulties, including those routinely experienced in adjustment to a new setting as well as more severe issues such as suicidality (Furr, Westefeld, McConnell, & Jenkins, 2001; Garlow et al., 2008; Haas, Hendin, & Mann, 2003).

4. Taxonomic Considerations

In the newest version of the Diagnostic and Statistical Manual (DSM-V; APA, 2013) the diagnostic criteria have changed for many disorders, including those assessed in the current study (Bögels et al., 2010; Greenberg, 2013). In some cases these modifications have been minimal, but in others changes have been more substantial and represent different theoretical approaches to clinical classification (Clark & Watson, 1991; Meehl, 2001; Widiger & Samuel, 2005; Kraemer et al., 2012). Future research applying methods similar to the current study will need to examine predictive validity in the context of this modified taxonomic system (Phillips et al., 2003; Andrews et al., 2010). Given the preliminary state of research progress validating the new DSM-V diagnostic criteria, lengthy time necessary for studies of this nature, and absence of updated diagnostic interview protocols it would be overly idealistic to expect this form of validation inquiry to occur in any appreciable volume in the near future (Greenberg, 2013). For

the time being diagnosis and outcome assessment must continue using the instruments that are available with a focus on practical dissemination until such a time that additional research supporting superior, more contemporary methods becomes available (Regier et al., 1998; Hatfield, 2004). Consideration of these issues is compounded by current movements in third-party payor systems and public policy to implement an international system of diagnostic classification (the International Classification of Diseases, 10th Edition, Clinician Modified; ICD-10 CM) as the professional standard; Miller & Luft, 1997; Phillips et al., 2003; Kupfer, Regier, & Kuhl, 2008). Additionally, the National Institute of Mental Health has publicly rejected adherence to the newest edition of the DSM on the basis of limited scientific support and shifting taxonomic standards, and has directed substantial funding to facilitating the development of general, integrated, transdiagnostic systems of classification (i.e., those focused on explication of endophenotypes; Clark & Watson, 1991; Kupfer, Regier, & Kuhl, 2008; Goldberg, 2014; Regier, Kaelber, Roper, Rae, & Sartorius, 1994). These current developments likely introduce difficulty in designing and conducting additional diagnostic prediction studies, given that the field is in a state of upheaval in determining which criteria and/or taxonomic system will be implemented in the near future (Piotrowski, 1999; Widiger & Samuel, 2005; Kraemer et al., 2012).

It is worth reiterating that diagnoses should not be assigned based on the results of screening instruments in isolation. Measures such as those used in the current study are simply tools that can be applied to improve the efficiency and accuracy of the diagnostic interview (Hatfield, 2004). The goals of their implementation may change via setting, and thus examination of their specific psychometric performance for a specific purpose becomes necessary to accumulate support for their usage (Glasgow at al., 2003). Additionally, best practice guidelines would indicate that scores on screening measures should always be coupled

with a more in-depth evaluation for the purpose of making definitive diagnoses (Zimmerman, 2010). Changes in scores should not necessarily be used to assess outcome improvement in instances in which the instrument utilized has not been empirically demonstrated to be able to detect incremental change (improvement or deterioration; Hudziak et al., 2004). That said, the index tests used in the current study have the advantage of being much briefer than the full "gold-standard" diagnostic process and being able to be completed autonomously by the respondent (Zimmerman, 1994). They are also easily accessible and freely available online (Beidas et al., 2014). Thus, additional examinations of dissemination and implementation of their use, contextualized in terms of setting, purpose, and specific psychometric support, will be necessary to continue to advance integrated assessment knowledge (Hunsley, 2003; Phillips et al., 2003; Jensen Doss, 2005). Although this burgeoning area of study may currently have significant limitations bridging to clinical practice, the alternative is too often to abandon structure in favor of more intuitive, idiographic assessment approaches (Riotrowski, 1999; Grove et al., 2000; Swets et al., 2000). This strategy is known to be unreliable and invalid (e.g., Dawes et al., 1989), and in the face of *any* alternative approaches supported by actuarial evidence represents limited infusion of science into practice (Piotrowski, 1999; Hatfield & Ogles, 2007). Thus, in conclusion, much more research is necessary to substantiate the utility of psychological self-report instrumentation and advance its integrated use in healthcare more broadly.

V. REFERENCES

Acklin, M. W. (1996). Personality assessment and managed care. *Journal of Personality Assessment, 66,* 194–201.

Addis, M. E. (2002). Methods for disseminating research products and increasing evidence-based practice: Promises, obstacles, and future directions. *Clinical Psychology: Science and Practice, 9,* 367-378.

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology, 68,* 331-339.

Allen, L. B., McHugh, R. K., & Barlow, D. H (2008). Emotional disorders: a unified protocol. In D. H. Barlow (ed). *Clinical handbook of psychological disorders* (4th ed.). New York, NY: Guilford Press.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.

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

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

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.

Anderson, D. A., & Paulosky, C. A. (2004). A survey of the use of assessment instruments by eating disorder professionals in clinical practice. *Eating and Weight Disorders, 9(*3), 238–241.

Andrews, G. (1996). Comorbidity in neurotic disorders: The similarities are more important than the differences. In R. M. Rapee (Ed.), *Current controversies in the anxiety disorders* (pp. 3–20). New York: Guilford Press.

Andrews, G., Hobbs, M. J., Borkovec, T. D., Beesdo, K., Craske, M. G., Heimberg, R. G., ... & Stanley, M. A. (2010). Generalized worry disorder: a review of DSM-IV generalized anxiety disorder and options for DSM-V. *Depression and Anxiety*, *27*(2), 134-147.

Anthony, N. (1968). The use of facts and cues in clinical judgments from interviews. J*ournal of Clinical Psychology, 24,* 37–39.

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment, 10,* 176–181.

Antony, M. M., Moras, K., Meadows, E. A., Di Nardo, P. A., Utech, J. E., & Barlow, D. H. (1994). The diagnostic significance of the functional impairment and subjective distress criterion: An illustration with the DSM-IH-R anxiety disorders. *Journal of Psychopathology and Behavioral Assessment, 16,* 253-263.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271–285.

Bagley, C., & Genuis, M. (1991). Psychology of computer use: Sexual abuse recalled: Evaluation of a computerized questionnaire in a population of young adult males. *Perceptual and Motor Skills, 72***,** 287-288.

Barlow, D. H., Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Allen, L. B., & May, J. T. E. (2010). *Unified protocol for transdiagnostic treatment of emotional disorders: Therapist guide*. Oxford University Press.

Basco, M. R., Bostic, J. Q., Davies, D., Rush, A. J., Witte, B., Hendrickse, W., et al. (2000). Methods to improve diagnostic accuracy in a community mental health setting. *American Journal of Psychiatry, 157*(10), 1599–1605.

Beck, A. T., Epstein, N., Brown, G. & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56,* 893-897.

Becker, D., & Lamb, S. (1994). Sex bias in the diagnosis of borderline personality disorder and posttraumatic stress disorder. *Professional Psychological Research Practices, 25,* 55–61.

Behar, E., Alcaine, O., Zuellig, A. R., & Borkovec, T. D. (2003). Screening for generalized anxiety disorder using the Penn State Worry Questionnaire: a receiver operating characteristic analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*(1), 25-43.

Berland, N. W., Thompson, J. K., & Linton, P. H. (1986). Correlation between the EAT-26 and the EAT-40, the Eating Disorders Inventory, and the Restrained Eating Inventory. *International Journal of Eating Disorders, 5*(3), 569-574.

Beidas et al., Free, Brief, and Validated: Standardized Instruments for Low-Resource Mental Health Settings, *Cognitive and Behavioral Practice* (2014),

http://dx.doi.org/10.1016/j.cbpra.2014.02.002

Bickman, L. (1999). Practice makes perfect and other myths about mental health services. *294*(15), 1898–1899.

Blais, M., & Norman, D. (1997). A psychometric evaluation of the DSM-IV personality disorder criteria. *Journal of Personality Disorders, 11*, 168–176.

Blashfield, R. K. (1990). Comorbidity and classification. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders* (pp. 61-82). Washington, DC: American Psychiatric Press.

Bögels, S. M., Alden, L., Beidel, D. C., Clark, L. A., Pine, D. S., Stein, M. B., & Voncken, M. (2010). Social anxiety disorder: questions and answers for the DSM-V. *Depression and Anxiety, 27*(2), 168-189.

Bogenschutz, M. P., Nurnberg, H. G. (2000). Theoretical and methodological issues in psychiatric comorbidity. *Harvard Review of Psychiatry, 8*, 18-24.

Borkovec, T. D., & Ruscio, A. M. (2001). Psychotherapy for generalized anxiety disorder. *Journal of Clinical Psychiatry, 62*, 37-45,

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., ... & Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*, *138*(1), W1-12.

Brammer, R. (2002). Effects of experience and training on diagnostic accuracy. *Psychological Assessment, 14,* 110–113.

Brehmer, B. (1980). In one word: Not from experience. *Psychologica, 45*, 223-241.

Brenner, D., & Howard, K. I. (1976). Clinical judgment as a function of experience and information. *Journal of Clinical Psychology, 32,* 721-728.

Brown, C. D., & Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures. *Chemometrics Intelligence Lab Systems, 80,* 24-38.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107,* 179-192

Brown, T. A., Chorpita, B. F., Korotitsch, W., & Barlow, D. H. (1997). Psychometric properties of the Depression Anxiety Stress Scales (DASS) in clinical samples. *Behaviour Research and Therapy, 35,* 79–89.

Brown, T. A., Di Nardo, P. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview for DSM-IV (ADIS-IV).* San Antonio: The Psychological Corporation.

Carver, C. S., & White, T. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67,* 319–333.

Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice, 33*(5), 446–453.

Cepoiu, M., McCusker, J., Cole, M. G., Sewitch, M., Belzile, E., & Ciampi, A. (2008). Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *Journal of General Internal Medicine, 23,* 25-36.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L., Calhoun, K. S., Crits-Christoph, P., et al. (1998). Update on empirically validated therapies, II. *Clinical Psychologist, 51*(1), 3–16.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Christoph, P., et al. (1996). An update on empirically validated therapies. *Clinical Psychologist, 49,* 5–18.

Chorpita, B. F., Bernstein, A., & Daleiden, E. L. (2008). Driving with roadmaps and dashboards: Using information resources to structure the decision models in service organizations. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(1-2), 114-123.

Chorpita, B. F., & Nakamura, B. J. (1998). Dynamic structure in diagnostic structured interviewing: A comparative test of accuracy and efficiency. *Journal of Psychopathology and Behavioral Assessment, 30,* 52-60.

Chorpita, B. F., Yim, L. M., Moffitt, C. E., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A Revised Child Anxiety and Depression Scale. *Behaviour Research and Therapy, 38,* 835–855.

Clara, I. P., Cox, B. J., & Enns, M. W. (2001). Confirmatory factor analysis of the Depression-Anxiety-Stress scales in depressed and anxious parents. *Journal of Psychopathology and Behavioral Assessment, 23,* 61–67.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100,* 316-336.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review Psychology, 46,* 121–53.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cook, L. J. (2007). Striving to help college students with mental health issues. *Journal of Psychosocial Nursing and Mental Health Services, 45*(4), 40-44.

Cook, T. D., & Campbell, T. D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Company.

Cook, C., Cleland, J., & Huijbregts, P. (2007). Creation and critique of studies of diagnostic accuracy: use of the STARD and QUADAS methodological quality assessment tools. *Journal of Manual & Manipulative Therapy*, *15*(2), 93-102.

Crawford, J. R., Garthwaite, P. H., Lawrie, C. J., Henry, J. D., MacDonald, M. A., Sutherland, J., & Sinha, P. (2009). A convenient method of obtaining percentile norms and accompanying interval estimates for self-report mood scales (DASS, DASS-21, HADS, PANAS, and SAD). *British Journal of Clinical Psychology*, *48*(2), 163-180.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. 16*, 297-334.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*(1), 75.

Dawes, M. (2005). Critically appraised topics and evidence-based medicine journals. *Singapore Medical Journal, 46,* 442–448.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351–367). Hillsdale, NJ: Erlbaum.

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ, 323*, 157-162.

Dempster, A. P., Laired, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

De Vet, H. C. W., Van der Weijden, T., Muris, J. W. M., Heyrman, J., Buntinx, F., & Knottnerus, J. A. (2001). Systematic reviews of diagnostic research. Considerations about assessment and incorporation of methodological quality. *European Journal of Epidemiology, 17*(4), 301-306.

Di Nardo, P. A., Brown, T. A., Lawton, J. K., & Barlow, D. H. (1995, November). *The Anxiety Disorders Interview Schedule-for DSM-IV Lifetime Version: Description and initial evidence for diagnostic reliability.* Paper presented at the meeting of the Association for Advancement of Behavior Therapy, Washington, DC.

Di Nardo, P. A., Moras, K., Barlow, D. H., Rapee, R. M., & Brown, T. A. (1993). Reliability of DSM-III-R anxiety disorder categories using the Anxiety Disorders Interview Schedule-Revised (ADIS-R). *Archives of General Psychiatry, 50,* 251-256.

Dotti, A., & Lazzari, R. (1998). Validation and reliability of the Italian EAT-26. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity, 3*(4), 188-194.

Edwards, A. (2003). Communicating risks: means that patients too have to learn to live with

uncertainty. *BMJ: British Medical Journal*, *327*(7417), 691.

Eisman, E., Dies, R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Meyer, G. J., & Moreland, K. L. (2000). Problems and limitations in the use of psychological assessment in contemporary health care delivery. *Professional Psychology: Research and Practice, 31*, 131-140.

Ellard, K. K., Fairholme, C. P., Boisseau, C. L., Farchione, T. J., & Barlow, D. H. (2010). Unified protocol for the transdiagnostic treatment of emotional disorders: Protocol development and initial outcome data. *Cognitive and Behavioral Practice*, *17*(1), 88-101.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY; The Guilford Press.

Fan, X. B., Thompson, & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling 6,* 56-83.

Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., ... & Barlow, D. H. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: a randomized controlled trial. *Behavior Therapy*, *43*(3), 666-678.

Fawcett, T. (2004). *ROC Graphs: Notes and practical considerations for data mining researchers*. Kluwer Academic Publishers, Netherlands.

Finn, S. E. (1982). Base rates, utilities, and DSM-III: Shortcomings of fixed-rule systems of psychodiagnosis. *Journal of Abnormal Psychology*, *91*, 294-302.

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Medicine*, *29,* 1043-1051.

Foa, E. B, Riggs, D. S., Dancu, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress, 6*(4), 459- 473.

Frances, A., Widiger, T., & Fyer, M. R. (1990). The influence of classification methods on comorbidity. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders* (pp. 41-59). Washington, DC: American Psychiatric Press.

Furr, S. R., Westefeld, J. S., McConnell, G. N., & Jenkins, J. M. (2001). Suicide and depression among college students: A decade later. *Professional Psychology: Research and Practice, 32*(1), 97.

Gadol, I. (1969). The incremental and predictive validity of the Rorschach test in personality assessments of normal, neurotic, and psychotic subjects. *Dissertation Abstracts, 29*, 3482-B.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, 105, 387-396.

Garb, H. (1994). Cognitive heuristics and biases in personality assessment. In L. Heath, L. Tindale, & J. Edwards, et al. (Eds.) *Applications of heuristics and biases to social issues* (pp. 73-90). New York, NY: Springer Press.

Garb, H. (1995). Sex bias and the diagnosis of borderline personality disorder. *Professional Psychology: Research and Practice, 26*, 526.

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research and Practice, 27*(3), 272-277.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice, 4*, 99-120.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment.* Washington, DC: American Psychological Association.

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15*, 508–520.

Garb, H. N. (2005). Clinical judgment and decision-making. *Annual Review of Clinical Psychology, 1*(1), 67–89.

Garlow, S. J., Rosenberg, J., Moore, J. D., Haas, A. P., Koestner, B., Hendin, H., & Nemeroff, C. B. (2008). Depression, desperation, and suicidal ideation in college students: results from the American Foundation for Suicide Prevention College Screening Project at Emory University. *Depression and anxiety, 25*(6), 482-488.

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: psychometric features and clinical correlates. *Psychological Medicine, 12*(04), 871-878.

Garvey, M. J., & Tuason, V. B. (1980). Mania misdiagnosed as schizophrenia. *Journal of Clinical Psychiatry, 41*, 75-78.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York, NY: Simon and Schuster.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ: British Medical Journal, 327*(7417), 741.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*(2), 53-96.

Gigerenzer, G., Wegwarth, O., & Feufel, M. (2010). Misleading communication of risk. *BMJ: British Medical Journal, 341*, 791-792.

Glasgow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy to effectiveness transition. *American Journal of Public Health, 93,* 1261–1267.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt test. *Journal of Consulting Psychology*, *23,* 25–33.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*, 483–496.

Goldberg, J. F. (2014). What should ICD-11 do that DSM-5 did not? *Australian and New Zealand Journal of Psychiatry*, *48*(1), 87-88.

Green, D. M., & Swets, J. A. (1996). *Signal detection theory and psychophysics.* New York: John Wiley & Sons, Inc.

Greenberg, G. (2013). *The book of woe: The DSM and the unmaking of psychiatry*. New York, NY: Penguin.

Griffith, L.F. (1997). Surviving no-frills mental health care: The future of psychological assessment. *Journal of Practical Psychiatry and Behavioral Health, 3,* 255–258.

Grigg, A. E. (1958). Experience of clinicians, and speech characteristics and statements of clients as variables in clinical judgment. *Journal of Consulting Psychology, 22*, 315–319.

Grove, W. M. (1987). The reliability of psychiatric diagnosis. In Last, C. G., Hersen, M. (eds), *Issues in Diagnostic Research* (pp. 99-119). New York, NY: Plenum.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Grove, W. M., Andreason, N. C., McDonald-Scott, P., Kelle, M. B., & Shaprio, R. W. (1981).

Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry, 38,* 408-413.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Grubaugh, A. L., Elhai, J. D., Cusack, K. J., Wells, C., & Frueh, B. C. (2007). Screening for PTSD in public-sector mental health settings: the diagnostic utility of the PTSD checklist. *Depression and Anxiety*, *24*(2), 124-129.

Haas, A. P., Hendin, H., & Mann, J. J. (2003). Suicide in college students. *American Behavioral Scientist, 46*(9), 1224-1240.

Hammond, K. R. (1996). How convergence of research paradigms can improve research on diagnostic judgment. *Medical Decision Making, 16,* 281–287.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839-843.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological bulletin*, *117*(1), 167.

Hatfield, D. R., & Ogles, B. M. (2004). The Use of Outcome Measures by Psychologists in Clinical Practice. *Professional Psychology: Research and Practice, 35*(5), 485–491.

Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research, 34*(3), 283–291.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*(4), 456.

Haynes, S. N., Nelson, K., & Blaine, D. D. (1999). Psychometric issues in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Ed). *Handbook of research methods in clinical psychology* (2nd ed., pp. 125-154). Hoboken, NJ: John Wiley & Sons.

Heimberg, R. G., Hope, D. A., Dodge, C. S., & Becker, R. E. (1990). DSM-III-R subtypes of social phobia: Comparison of generalized social phobics and public speaking phobics. *The Journal of nervous and mental disease*, *178*(3), 172-179.

Henry, J. D., & Crawford, J. R. (2005). The 21-item version of the Depression Anxiety Stress Scales (DASS–21): Normative data and psychometric evaluation in a large non-clinical sample. *British Journal of Clinical Psychology*, *44*, 227-239.

Hudson, J. I., Hiripi, E., Pope Jr, H. G., & Kessler, R. C. (2007). The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. *Biological Psychiatry*, *61*(3), 348-358.

Hudziak, J. J., Copeland, W., Stanger, C., & Wadsworth, M. (2004). Screening for DSM-IV externalizing disorders with the Child Behavior Checklist: a receiver-operating characteristic analysis. *Journal of Child Psychology and Psychiatry, 45*(7), 1299-1307.

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, *15*(4), 443.

Hyman, S. E. (2003). Diagnosing disorders. *Scientific American, 289*(3), 96–103.

Jensen Doss, A. (2005). Evidence-based diagnosis: Incorporating diagnostic instruments into clinical practice. *Journal of the American Academy of Child & Adolescent Psychiatry. 44*(9), 947–952.

Kadison, R., & DiGeronimo, T. F. (2004). *College of the overwhelmed: The campus mental*

health crisis and what to do about it. San Francisco, CA: Jossey-Bass.

Kendall-Taylor, N., & Mikulak, A. (2009). *Child mental health: A review of the scientific discourse.* Washington, D.C: Frameworks.

Kendler, K. S., Karkowski, L. M., & Prescott, C. A. (1999). Fears and phobias: reliability and heritability. *Psychological Medicine, 29*(03), 539-553.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395-399.

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

Kessler, R. C., Berglund, P. A., Demler, O., Jin, R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of General Psychiatry, 62,* 593-602.

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

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

Kessler, R. C., & Üstün, T. B. (2004). The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). *International Journal of Methods in Psychiatric Research*, *13*(2), 93-121.

Kessler, R. C., Wittchen, H. U., Abelson, J. M., Mcgonagle, K., Schwarz, N., Kendler, K. S., ... & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US national comorbidity survey (NCS). *International Journal of Methods in Psychiatric Research*, *7*(1), 33-55.

Kitzrow, M. A. (2003). The mental health needs of today's college students: Challenges and recommendations. *NASPA Journal*, *41*(1), 165-179.

Koenigsberg, H. W, Kaplan, R. D., Gilmore, M. M, Cooper, A. M. (1985). The relationship between syndrome and personality disorder in DSM-III: experience with 2,462 patients. *American Journal of Psychiatry, 142,* 207-212.

Kraemer, H. C. (1988). Assessment of 2 x 2 associations: Generalizations of signal-detection methodology. *American Statistician, 42,* 37-49.

Kraemer, H. C. *Evaluating medical tests: Objective and quantitative guidelines.* Newbury Park: Sage Publications, 1992.

Kraemer, H. C. (2003). Current concepts of risk in psychiatric disorders. *Current Opinion in Psychiatry*, *16*(4), 421-430.

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science, 20* (3), 223-230.

Kraemer, H. C. (2013). Toward Sound Objective Evaluation of Clinical Measures. *The American Journal of Geriatric Psychiatry*, *21*(7), 589-595.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry, 59,* 990–996.

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). DSM-5: how reliable is reliable enough?. *American Journal of Psychiatry*, *169*(1), 13-15.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(12), 1524-1529.

Kroenke, K., Spitzer, R. L., Williams, J. B., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of internal medicine*, *146*(5), 317-325.

Kupfer, D. J., Regier, D. A., & Kuhl, E. A. (2008). On the road to DSM-V and ICD-11. *European Archives of Psychiatry and Clinical Neuroscience*, *258*(5), 2-6.

Lambert, D., & Hartley, D. (1998). Linking primary care and rural psychiatry: where have we been and where are we going?. *Psychiatric Services*, *49*(7), 965-967.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D.A., Nielsen, S. L., Smart, D. W. (2003). Is It Time for Clinicians to Routinely Track Patient Outcome? A Meta-Analysis. *Clinical Psychology: Science and Practice*.*10*(3), 288-301.

Lang, A. J., & Stein, M. B. (2005). An abbreviated PTSD checklist for use as a screening instrument in primary care. *Behaviour Research and Therapy*, *43*(5), 585-594.

Lilienfeld, S., Ruscio, J., & Lynn, S. J. (2008) Navigating the mindfield: A user's guide to separating science from pseudoscience in mental health. *Amherst, NY: Prometheus Books.*

Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an

introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation of Clinical Practice, 12,* 132–139.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Little, R. J. A., & Rubin, D. B. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202.

Loring, M., & Powell, B. (1988). Gender, race and DSM-III: a study of the objectivity of psychiatric diagnostic behavior. *Journal of Health and Social Behavior*, *29,* 1–22.

Lovibond, S. H., & Lovibond, P. F. (1993). *Manual for the Depression Anxiety Stress Scales (DASS): Psychology Foundation Monograph*. Wales, Australia: University of New South Wales Psychology Foundation.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy, 33*, 335-342.

Loy, C. T, & Irwig, L. (2004). Accuracy of diagnostic tests read with and without clinical information: a systematic review. *Journal of the American Medical Association, 292*(13), 1602-1609.

Löwe, B., Spitzer, R. L., Williams, J. B., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *General hospital psychiatry*, *30*(3), 191-199.

Luft, J. (1950). Implicit hypotheses and clinical predictions. *Journal of Abnormal and Social Psychology, 45,* 756–760.

Mackinnon, A., & Mulligan, R. (1998). Combining cognitive testing and informant report to increase accuracy in screening for dementia. *American Journal of Psychiatry, 155*, 1529-1535.

Martin, M., Horder, P., & Jones, G. V. (1992). Integral bias in naming of phobia-related words. *Cognition & Emotion, 6*(6), 479-486.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215–241.

Meehl P. (2001). Comorbidity and taxometrics. *Clinical Psychology Science and Practice, 8,* 507–519.

Meehl, P. E. (1954). Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

McHugh, R. K., Daughters, S. B., Lejuez, C. W., Murray, H. W., Hearon, B. A., Gorka, S. M., & Otto, M. W. (2011). Shared variance among self-report and behavioral measures of distress intolerance. *Cognitive Therapy and Research, 35*(3), 266-275.

Mellor-Clark, J., Barkham, M., Connell, J. & Evans, C. (1999). Practice-based evidence and standardized evaluation: Informing the design of the CORE system. *European Journal of Psychotherapy, Counselling and Health, 2*, 357-374

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminal Nuclear Med, **8***, 283–298.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy, 28,* 487-195.

Mezzich, J. E, Fabrega, H., Coffman, G. A., & Haley R. (1989). *DSM-III* disorders in a large sample of psychiatric patients: frequency and specificity of diagnoses. *American Journal of Psychiatry, 14*(6), 212- 219

Miller, P. R. (2001). Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews. *Psychiatry Research, 105,* 265–271.

Miller, P. R., Dasher, R., & Collins, R., (2001). Inpatient diagnostic assessments, 1: accuracy of structured versus unstructured interviews. *Psychiatry Research, 105,* 255-264.

Miller, R. H., & Luft, H. S. (1997). Does managed care lead to better or worse quality of care? *Health Affairs, 16,* 7–25.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet, 374,* 609– 619.

Mojtabai, R., & Nicholson, R. A. (1995). Interrater reliability of ratings of delusions and bizarre delusions. *American Journal of Psychiatry, 152,* 1804–1806.

Mowbray, C. T., Megivern, D., Mandiberg, J. M., Stein, C. H., Collins, K., Sandra Kopels, J. D., et al. (2006). Campus Mental Health Services: Recommendations for Change. *American Journal of Orthopsychiatry, 76*(2), 226-237.

Mulvihill, M. E. (2005). Karraker v. Rent-A-Center: Testing the Limits of the ADA, Personality Tests, and Employer Preemployment Screening. *Loy. U. Chi. LJ*, *37*, 865.

Muris, P., Mayer, B., Huijding, J., & Konings, T. (2008). A dirty animal is a scary animal! Effects of disgust-related information on fear beliefs in children. *Behaviour research and therapy*, *46*(1), 137-144.

Muris, P., Merckelbach, H., de Jong, P. J., & Ollendick, T. H. (2002). The etiology of specific fears and phobias in children: A critique of the non-associative account. *Behaviour Research and Therapy*, *40*(2), 185-195.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic tests: application of receiver operating characteristics analysis. *Archives of General Psychiatry, 44,* 550- 555

Myers, H. F., Lesser, I., Rodriguez, N., Mira, C. B., Hwang, W. C., Camp, C., et al. (2002). Ethnic differences in clinical presentation of depression in adult women. *Cultural Diversity and Ethnic Minority Psychology, 8,* 138-156.

Ng, F., Trauer, T., Dodd, S., Callaly, T., Campbell, S., & Berk, M. (2007). The validity of the 21-item version of the Depression Anxiety Stress Scales as a routine clinical outcome measure. *Acta Neuropsychiatrica*, *19*(5), 304-310.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods, 5*(2), 241.

Norton, P. J. (2007). Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups. *Anxiety, stress, and coping*, *20*(3), 253-265.

Nunnally, J. C. (1978). *Psychometric Theory.* New York, NY: McGraw-Hill.

O'Boyle, M., & Self, D. (1990). A comparison of two interviews for DSM-III-R personality

disorders. *Psychiatry Reserves, 32*,85-92.

O¨st, L.-G. (1985). Ways of acquiring phobias and outcome of behavioural treatment. *Behaviour Research and Therapy, 23,* 683– 689.

Oldham, J. M., & Skodol, A. E. (1991). Personality disorders in the public sector. *Hospital and Community Psychiatry, 42,* 481–487.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76*(28), Whole No. 547).

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29,* 261–265.

Osman, A., Wong, J. L., Bagge, C. L., Freedenthal, S., Gutierrez, P. M., & Lozano, G. (2012). The Depression Anxiety Stress Scales—21 (DASS-21): Further Examination of Dimensions, Scale Reliability, and Correlates. *Journal of clinical psychology*, *68*(12), 1322-1338.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

Pereira-Maxwell, F. (1998). *A-Z of medical statistics: a companion for critical appraisal.* London, England: Arnold.

Phelps, R., Eisman, E. J., & Kohout, J. (1998). Psychological practice and managed care: Results of the CAPP practitioner survey. *Professional Psychology: Research and Practice, 29,* 31–36.

Phillips, K. A., First, M. B., & Pincus, H. A. (2003). *Advancing DSM: Dilemmas in psychiatric diagnosis.* Washington, DC: American Psychiatric Association.

Pincus, H. A., Zarin, D. A., Tanielian, T. L., Johnson, J. L., West, J. C., Pettit, A. R., Marcus, S. C., Kessler, S. C., & McIntyre, J. (1999). Psychiatric patients and treatments in 1997: Findings from the American Psychiatric Association Practice Research Network (1999). *Archives General Psychiatry, 56*, 441-449.

Piotrowski, C. (1992a). Assessment training and practice in professional psychology: Data, issues, and perspectives [Special issue]. *The Journal of Training & Practice in Professional Psychology, 6*(1).

Piotrowski, C. (1992b). Assessment training and practice in professional psychology II: Commentaries and reviews [Special issue]. *The Journal of Training & Practice in Professional Psychology, 6*(2).

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55*, 787-796.

Piotrowski, C., & Keller, J.W. (1992). Psychological testing in applied settings. A literature review from 1982–1992. *Journal of Training & Practice in Professional Psychology, 6*(2), 74–82.

Piotrowski, C., Belter, R. W., & Keller, J. W. (1998). The impact of "managed care" on the practice of psychological testing: Preliminary findings. *Journal of Personality Assessment, 70*, 441-447.

Purdon, C. (1999). Thought suppression and psychopathology. *Behaviour Research and Therapy*, *37*(11), 1029-1054.

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C., &

Norquist, G. S. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders: Implications for research and policy. *Archives of General Psychiatry, 55*, 109-115.

Regier, D. A., Kaelber, C. T., Roper, M. T., Rae, D. S., & Sartorius, N. (1994). The ICD-10 clinical field trial for mental and behavioral disorders: results in Canada and the United States. *American Journal of Psychiatry, 151*(9), 1340-1350.

Reid, M. C., Lachs, M. S., & Feinstein, A. R. (1995). Use of methodological standards in diagnostic test research: Getting better but still not good. *JAMA, 274*(8), 645-651.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184.

Robins, L. N., & Regier, D. A. (1991). *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. New York, NY: Free Press.

Roemer, L., & Borkovec, T. D. (1994). Effects of suppressing thoughts about emotional material. *Journal of Abnormal Psychology, 10*, 467–474.

Rogers, E. M. (2004). A prospective and retrospective look at the diffusion model. *Journal of Health Communication, 9*(S1), 13-19.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Rogers, F, S. L., Gordon, C. Y., Schanzenbacher, K. E., & Case-Smith, J. (2001). Common diagnosis in pediatric occupational therapy practice. In J. Case-Smith (Ed.), *Occupational therapy for children* (4th ed., pp. 136–187). St. Louis, MO: Mosby-Year Book.

Ruscio, A. M., Brown, T. A., Chiu, W. T., Sareen, J., Stein, M. B., & Kessler, R. C. (2008).

Social fears and social phobia in the USA: results from the National Comorbidity Survey Replication. *Psychological Medicine, 38*(01), 15-28.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London, England: Chapman & Hall.

Schafer, J.L. & Olsden, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545-571.

Segal, D. L., Corcoran, J., & Coughlin, A. (2002). Diagnosis, differential diagnosis, and the SCID. In M. Hersen & L. K. Porzelius (Eds.), *Diagnosis, conceptualization, and treatment planning for adults: A step-by-step guide* (pp. 13–34). Mahwah, NJ: Erlbaum.

Selman, T. J., Morris, R. K., Zamora, J., & Khan, K. S. (2011). The quality of reporting of primary test accuracy studies in obstetrics and gynecology: Application of the STARD criteria. *BMC Women's Health, 11*(1), 1-7.

Shear, M. K., Greeno, C., Kang, J., et al (2000). Diagnosis of nonpsychotic in community clinics. *American Journal of Psychiatry, 157,* 581-587.

Siervo, M., Boschi, V., Papa, A., Bellini, O., & Falconi, C. (2005). Application of the SCOFF, Eating Attitude Test 26 (EAT 26) and Eating Inventory (TFEQ) Questionnaires in young women seeking diet-therapy. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity, 10*(2), 76-82.

Signal Detection or ROC Software: Version 5. (2007). http://mirecc.stanford.edu

Silverman, G. (1969). Manic depressive psychosis. *British Journal of Psychiatry, 115*(520), 371-372.

Schmidt, N. B., Richey, J. A., & Fitzpatrick, K. K. (2006). Discomfort intolerance: Development

of a construct and measure relevant to panic disorder. *Journal of Anxiety Disorders*, *20*(3), 263-280.

Schulberg, H. C., Saul, M., McClelland, M., Ganguli, M., Christy, W., & Frank, R. (1985). Assessing depression in primary medical and psychiatric practices. *Archives of General Psychiatry*, *42*(12), 1164-1170.

Spitzer, R. L., & First, M. B. (2005). Classification of psychiatric disorders. *JAMA, 294*(15), 1898-1899.

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Journal of the American Medical Association, 282*, 1737–1744.

Spitzer, R.L., & Williams, J. B. (1994). American psychiatry's transformation following the publication of DSM-III. *American Journal of Psychiatry, 151*(3), 459–460.

Stanley, M. A., Diefenbach, G. J., Bourland, S. L., Rodriguez, H, & Wagener, P. (2003). Cognitive-Behavior Therapy for Late-Life Generalized Anxiety Disorder in Primary Care: Preliminary Findings. *American Journal of Geriatric Psychiatry, 11(*1), 92-96.

Stout, C. E., & Cook, L. P. (1999). New areas for psychological assessment in general health care settings: What to do today to prepare for tomorrow. *Journal of Clinical Psychology, 55*, 797–812.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52*, 121–128.

Streiner, D., & Norman, G. (2003). *Health Measurement Scales: A practical guide to their development and use* (3rd ed.). New York, NY: Oxford University Press.

Streiner, D.L., & Norman, G. R. (2006). "Precision" and "accuracy": Two terms that are neither.

Journal Clinical Epidemiology, 59, 327–330.

Summerfeldt, L. J., & Antony, M. M. (2002). Structured and semistructured diagnostic interviews. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment, treatment planning, and outcome evaluation: Empirically supported strategies for psychological disorders* (pp. 3–37). New York: Guilford Press.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Swets, J. A., R. M. Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283,* 82-87.

Swets, J.A., & Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press: New York.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson Education Incorporated.

Taylor, S., & Cox, J. (1998). An Expanded Anxiety Sensitivity Index: Evidence for a Hierarchic Structure in a Clinical Sample. *Journal of Anxiety Disorders, 12*(5), 463-483.

Taylor, S., Zvolensky, M. J., Cox, B. J. Deacon, B., Heimberg, Richard, G., et al. (2007). Robust dimensions of anxiety sensitivity: Development and initial validation of the Anxiety Sensitivity Index-3. *Psychological Assessment, 19*(2), 176-188.

Trent, L. R., Buchanan, E., Ebesutani, C., Ale, C., Heiden, L., Hight, T. et al. (2012). A Measurement Invariance Examination of the Revised Child Anxiety and Depression Scale in a Southern Sample: Differential Item Functioning Between African American and Caucasian Youth. *Assessment, 19*(3). 1-13.

Tripepi, G., Jager, K., J., Dekker, F. W., & Zoccali, C. (2009). Diagnostic methods 2: receiver operating characteristic (ROC) curves. *Kidney International, 76,* 252-256.

Turner, D. R. (1966). Predictive efficiency as a function of amount of information and level of professional experience. *Journal of Projective Techniques and Personality Assessment, 30*, 4-11.

Van Ryn, M., & Burke, J. (2000). The effect of patient race and socioeconomic status on physicians' perceptions of patients. *Social Sciences and Medicine, 50,* 813–*28.*

Voelker, R. (2003). Mounting student depression taxing campus mental health services. *JAMA, 289*(16), 2055-2056.

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., & Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Annals of Internal Medicine, 156*(5), 340-349.

Weisz, J. R,, & Addis, M. E. (2006). The research-practice tango and other choreographic challenges: using and testing evidence-based psychotherapies in clinical i.e., STARD criteria; Bossuyt, et al., 2003; Kraemer, 1992) care settings. In C. D. Goodheart, A. E. Kazdin, & R. J. Sternberg (Eds.). *Evidence-based psychotherapy: Where practice and research meet* (pp. 179-206). Washington, DC: American Psychological Association.

Wetzler, S., & van Praag, H. M. (1989). Assessment of depression. In S. Wetzler (Ed.), *Measuring Mental Illness: Psychometric assessment for clinicians* (pp. 71-88). Washington, DC: American Psychiatric Association.

Whiting, P., Rutjes, A. W., Dinnes, J., Reitsma, J., Bossuyt, P. M., & Kleijnen, J. (2004). Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technology Assessment (Winchester, England), 8*(25), iii-1.

Widiger, T. A, & Sankis, L. M. (2000). Adult psychopathology: issues and controversies. *Annual Review of Psychology, 51,* 377-404.

Widiger, T. A., & Ford-Black, M. M. (1994). Diagnoses and disorders. *Clinical Psychology Science and Practice, 1*(1), 84–87.

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition. *Journal of Abnormal Psychology*, *114,* 494 -504.

Widiger, T. A., & Spitzer, R. L. (1991). Sex bias in the diagnosis of personality disorders: Conceptual and methodological issues. *Clinical Psychology Review, 11*, 1-22.

Wilczynski, N. L. (2008). Quality of reporting of diagnostic accuracy studies: No change since STARD statement publication—Before-and-after study 1. *Radiology*, *248*(3), 817-823.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594.

Williams, J. W., Noël, P., Cordes, J. A., Ramirez, G., & Pignone, M. (2002). Is This Patient Clinically Depressed? *JAMA*, *287*(9), 1160-1170.

Wolpe, J., & Lang, P. J. (1964). A fear survey schedule for use in behaviour therapy. *Behaviour Research and Therapy, 2,* 27–30.

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. (2002). Clinical assessment. *Annual Review of Psychology*, *53*, 519–543.

Woody, S. R., & Teachman, B. A. (2000). Intersection of disgust and fear: Normative and pathological views. *Clinical Psychology: Science and Practice*, *7*(3), 291-311.

World Health Organization. (1992). The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15,* 478-495.

Zhou, X., Obuchowski, N., & McClish D. (2002). *Statistical methods in diagnostic medicine* (pp. 171-176). New York, NY: John Wiley and Sons.

Zimmerman, M. (2003). Integrating the assessment methods of researchers into routine practice: the Rhode Island Methods to Improve Diagnostic and Assessment Services (MIDAS) project. In M. First (Ed.), *Standardized Evaluation in Clinical Practice* (pp. 29-74). Washington, DC: American Psychiatric Publishing Inc.

Zimmerman, M., & Mattia, J. I. (1999). Psychiatric diagnosis in clinical practice: Is comorbidity being missed? *Comprehensive Psychology, 40,* 182-191.

Zimmerman, M., & Mattia, J. I. (1999). The reliability and validity of a screening questionnaire for 13 DSM–IV Axis I disorders (the Psychiatric Diagnostic Screening Questionnaire) in psychiatric outpatients. *Journal of Clinical Psychiatry, 60,* 677– 683.

Zimmerman, M., & Sheeran, T. (2003). Screening for principal versus comorbid conditions in psychiatric outpatients with the Psychiatric Screening Questionnaire. *Psychological Assessment, 15,* 110–114.

Zimmerman, M., Farber, N. J., Hartung, J., Lush, D. T., & Kuzma, M. A. (1994). Screening for psychiatric disorders in medical patients: a feasibility and patient acceptance study. *Medical Care, 32,* 603- 608.

VI. APPENDIX

Table 1
Sample Characteristics

Age	*n*	%
18	82	28.1
19	114	39.0
20	28	9.6
21	28	9.6
22	15	5.1
23	7	2.4
≥ 24	16	5.5
Gender	***n***	**%**
Male	67	22.9
Female	226	77.1
Ethnicity	***n***	**%**
White	190	65.1
Black	83	28.4
Asian	8	2.7
Multi/Other	10	3.4
Diagnostic BR[1]	***n***	**%**
≥ 1 Disorder	193	66.1
≥ 2 Disorders	110	37.7
MDD	19	6.5
GAD	54	18.4
OCD	5	1.7
PD	21	7.2
PTSD	14	4.8
SAD	119	40.7
SP	121	41.4
ED[2]	35	12.0

[1]Diagnostic base-rates
[2]Anorexia nervosa (n=4), bulimia nervosa (n=14), and eating disorder not otherwise specified (n=17) were collapsed into a single “eating disorder” category (n=35).

Table 2
Intercorrelation Among Index Tests[8 9]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. AAQ																		
2. ASI-3	.460																	
3. BIS	.377	.273																
4. BAS	-.033[b]	-.007[b]	-.014[b]															
5. DASS-TOT	.424	.535	.295	-.056														
6. DASS-D	.426	.482	.236	-.069	.821													
7. DASS-A	.381	.452	.281	-.035	.905	.648												
8. DASS-S	.305	.500	.212	-.067	.819	.554	.614											
9. DERS	.581	.515	.355	-.031	.454	.491	.413	.288										
10. DIS	.024[b]	.243	-.032[b]	.206	.070[b]	.081[b]	.048[b]	.066[b]	.046[b]									
11. EAT-26	.174	.253	.207	.221	.275	.234	.242	.213	.187	.118[a]								
12. FSS	.435	.540	.389	-.051	.361	.274	.358	.315	.363	.078[b]	.373							
13. IUS	.537	.664	.440	-.071	.538	.481	.487	.420	.565	.179	.251	.578						
14. OCI-R	.252	.533	.221	-.009	.395	.302	.370	.349	.347	.298	.326	.560	.534					
15. PCL	.268	.358	.174[a]	.057	.432	.364	.372	.359	.313	.167[a]	.293	.408	.319	.459				
16. PSWQ	.483	.410	.609	-.155[a]	.405	.315	.375	.329	.413	.064[b]	.285	.400	.542	.380	.344			
17. SIAS	.482	.526	.489	-.149[a]	.380	.390	.335	.325	.443	.136[b]	.259	.591	.562	.463	.319	.341		
18. SPS	.409	.577	.410	-.162[a]	.348	.318	.309	.319	.378	.129[b]	.263	.644	.600	.537	.345	.412	.744	
19. WBSI	.445	.421	.225	.012[b]	.496	.466	.421	.409	.381	.064[b]	.143[a]	.305	.477	.377	.329	.346	.243	.298

Note. Sample sizes differed across measures. See Table 3 for corresponding sample sizes and explanation of inter-measure discrepancy (in footnote).
All correlations significant at the $p < .01$ level (2-tailed) unless otherwise denoted (see below).
[a] Significant at the $p < .05$ level (2-tailed).
[b] Not significant.
AAQ=Acceptance and Action Questionnaire; ASI-3= Anxiety Sensitivity Index-3; BIS= Behavioral Inhibition Scale; BAS= Behavioral Activation Scale; DASS-TOT= Depression Anxiety Stress Scale- Total Score; DASS-D= Depression Anxiety Stress Scale- Depression Subscale; DASS-A= Depression Anxiety Stress Scale- Anxiety Subscale; DASS-S= Depression Anxiety Stress Scale- Stress Subscale; DERS= Difficulties in Emotion Regulation Scale; DIS= Discomfort Intolerance Scale; EAT-26= Eating Attitudes Test-26; FSS= Fear Survey Schedule – III; IUS= Intolerance of Uncertainty Scale; OCI-R= Obsessive–Compulsive Inventory-Revised; PCL= PTSD Checklist; PSWQ= Penn State Worry Questionnaire; SPS= The Social Phobia Scale; SIAS= Social Interaction Anxiety Scale; WBSI= White Bear Suppression Inventory.

Table 3
Psychometric and Descriptive Data for Index Tests

Measure[1]	Abbrev[2]	Items[3]	Mean[4]	SD[5]	Alpha[6]	N[7]	Missing[8]	Reference[9]
Acceptance and Action Questionnaire	*AAQ*	9	33.16	6.68	.518	197	.334	Hayes, et al. (2004)
Anxiety Sensitivity Index-3	*ASI-3*	18	12.24	1.05	.883	292	.133	Taylor et al. (2007)
Behavioral Inhibition Scale	*BIS*	7	2.16	3.90	.755	292	.490	Carver & White (1994)
Behavioral Activation Scale	*BAS*	13	4.15	5.47	.815	291	.650	*(Same as above)*
Depression Anxiety Stress Scale: Total Score	*DASS-TOT*	21	1.39	9.20	.804	292	.817	Lovibond & Lovibond (1995)
Depression Anxiety Stress Scale: Depression Subscale	*DASS-D*	7	2.77	3.51	.855	292	.782	*(Same as above)*
Depression Anxiety Stress Scale: Anxiety Subscale	*DASS-A*	7	2.66	3.08	.738	292	.489	*(Same as above)*
Depression Anxiety Stress Scale: Stress Subscale	*DASS-S*	7	4.93	4.08	.820	292	.880	*(Same as above)*
Difficulties in Emotion Regulation Scale	*DERS*	36	4.28	16.84	.755	292	.152	Gratz & Roemer (2004)
Discomfort Intolerance Scale	*DIS*	5	13.92	4.81	.466	291	.137	Schmidt, Ritchey, & Fitzpatrick (2006)
Eating Attitudes Test-26	*EAT-26*	26	8.77	13.05	.287	291	.421	Garner et al., (1982)
Fear Survey Schedule-III	***FSS*	73	74.19	52.02	.973	197	.894	Wolpe & Lang (1964)
Intolerance of Uncertainty Scale	*IUS*	27	51.65	18.02	.915	292	.076	Buhr & Dugas (2002)
Obsessive–Compulsive Inventory-Revised	***OCI-R*	18	12.99	12.49	.919	292	.251	Foa et al. (2002)
PTSD Symptom Checklist	***PCL*	17	28.37	12.66	.941	194	.157	Weathers, Litz, Herman, Huska, & Keane (1993)
Penn State Worry Questionnaire	***PSWQ*	16	39.61	9.86	.788	194	.500	Meyer, Miller, Metzger, & Borkovec (1990)
Social Interaction Anxiety Scale	***SIAS*	20	22.47	13.14	.894	198	.400	Mattick & Clarke (1989)
Social Phobia Scale	***SPS*	20	13.62	13.58	.934	198	.200	*(Same as above)*
White Bear Suppression Inventory	*WBSI*	15	44.31	12.78	.917	196	.593	Wegner & Zanakos (1994)

Note. *Measure was added to protocol after study enrollment began resulting in a maximum sample size of N=198 participants. No imputation used for pre-existing cases and data were analyzed separately. Maximum sample size for all other measures (that comprised the initial battery) is N=292.

[1]Full name of measure

[2]Abbreviation used to refer to the corresponding measure throughout tables and text

[3]Number of items on respective scale

[4]Mean score

[5]Standard deviation of mean score

[6]Reliability coefficient (i.e., internal consistency measure; Cronbach's alpha [Cronbach, 1951])

[7]N represents number of participants used in all analyses for corresponding measure; Maximum N=292 unless otherwise noted (*****) in *Abbrev* (i.e., Abbreviation) column (See *Note* above)

[8]Percent of missing data replaced

[9]Development article reference

Table 4

Comparison of Index Test Scores Within Diagnostic Categories

	Mean Score of Positive (+) versus Negative (−) Diagnostic Groups and Effect Size (Cohen's *d*)																										
	≥1			≥2			MDD			GAD			PD			PTSD			SAD			SP			ED		
	$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$			$\bar{x}$		
	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*	+	−	*d*
1	25.4	21.8	0.54	26.5	22.8	0.58	27.3	23.9	0.51	27.6	23.4	0.66	28.9	23.8	0.78	27.6	24	0.55	25.9	23.0	0.45	25.4	23.3	0.32	23.5	24.3	1.60
2	14.2	8.6	0.58	15.3	10.5	0.49	19.2	11.8	0.75	16.3	11.4	0.49	18.9	11.8	0.72	17.4	12.1	0.53	15.4	10.2	0.53	14.5	10.7	0.38	23.5	24.3	-0.10
3	25.9	27.5	-0.29	26.1	26.6	-0.09	24.5	26.6	-0.38	24.6	26.9	-0.43	25.1	26.5	-0.27	27.4	26.4	0.18	25.7	27.0	-0.24	26.7	26.2	0.09	13.6	12.1	0.14
4	13.8	12.1	0.42	14.1	12.7	0.38	13.5	13.2	0.09	13.8	13.1	0.20	15.0	13.1	0.50	14.9	13.1	0.45	14.4	12.4	0.54	13.7	12.9	0.20	25.9	26.5	-0.10
5	3.2	1.6	0.55	3.6	2.1	0.48	6.5	2.5	1.16	5.2	2.2	0.89	4.9	2.6	0.64	6.4	2.6	1.09	3.4	2.4	0.28	3.1	2.5	0.17	13.5	13.2	0.10
6	3.4	1.6	0.52	3.8	2.2	0.49	4.3	2.6	0.58	4.4	2.3	0.73	5.1	2.5	0.88	4.2	2.6	0.53	3.4	2.2	0.39	3.4	2.2	0.40	3.2	2.7	0.14
7	5.6	3.6	0.51	6.2	4.2	0.51	8.2	4.7	0.87	8.1	4.2	1.00	7.2	4.8	0.61	7.8	4.8	0.74	5.8	4.3	0.37	5.7	4.4	0.31	3.5	2.6	0.32
8	12.2	6.8	0.62	13.6	8.5	0.58	19.0	9.8	1.03	17.7	8.7	1.06	17.2	9.9	0.82	18.4	10.0	0.93	12.6	8.9	0.41	12.2	9.1	0.34	5.0	4.9	0.02
9	42.9	35	0.48	45.2	37.3	0.48	41.1	40.2	0.05	46.4	38.9	0.45	50.5	39.5	0.66	50.7	39.7	0.65	44.1	37.6	0.38	43.5	38.0	0.33	11.8	10.2	0.17
10	14.1	13.6	0.10	14.2	13.8	0.07	14.5	13.9	0.13	14.8	13.7	0.22	17.0	13.7	0.69	14.2	13.9	0.06	13.8	14.0	-0.06	14.2	13.8	0.09	43.7	39.8	0.23
11	9.0	8.3	0.05	9.6	8.3	0.10	8.2	8.8	-0.05	9.5	8.6	0.06	11.0	8.6	0.18	10.6	8.7	0.14	8.6	8.9	-0.03	9.7	8.1	0.12	14.2	13.9	0.06
12	85.0	56.6	0.56	90.7	67.5	0.45	80.1	75.1	0.09	92.5	71.9	0.40	105.8	73.3	0.63	68.3	75.9	-0.15	96.9	61.6	0.72	90.3	65.7	0.48	10.6	8.5	0.16
13	28.3	17.5	0.62	30.7	21.0	0.56	33.9	24.0	0.55	34.9	22.3	0.73	39.4	23.5	0.90	31.9	24.3	0.42	30.5	20.6	0.57	28.8	21.7	0.40	83.1	75.0	0.15
14	14.9	9.7	0.42	16.4	11.5	0.40	15.8	12.9	0.23	16.6	12.4	0.33	21.9	12.5	0.76	12.6	13.2	-0.05	15.9	11.3	0.37	17.0	10.6	0.53	25.5	24.5	0.05
15	13.8	7.3	0.53	17.6	8.5	0.77	20.4	10.9	0.76	17.4	1.4	0.56	23.1	10.8	1.00	31.4	10.4	1.80	14.8	9.5	0.43	15.9	8.7	0.60	15.9	13.0	0.23
16	33.0	25.7	0.57	36.9	27.3	0.77	38.5	29.9	0.65	38.6	28.8	0.76	38.0	30.0	0.61	35.8	30.2	0.42	34.8	27.7	0.55	34.1	28.2	0.45	17.3	11.3	0.48
17	26.2	15.6	0.88	25.5	21.1	0.34	20.8	22.7	-0.15	27.9	21.5	0.49	30.7	22.0	0.67	23.3	22.5	0.06	28.7	18.7	0.82	25.0	21.0	0.31	33.9	30.3	0.27
18	17.2	7.2	0.78	19.3	11.0	0.63	14.4	13.7	0.05	19.3	12.6	0.49	26.9	12.9	1.06	16.5	13.6	0.21	20.4	9.5	0.86	17.6	11.3	0.47	26.2	22.4	0.29
19	31.9	25.3	0.53	33.4	27.3	0.49	38.1	29.0	0.72	35.7	28.2	0.61	38.4	28.9	0.76	38.1	29.2	0.71	32.0	28.0	0.33	32.1	27.9	0.33	21.8	13.3	0.63

Note. See Table 1 for corresponding sample sizes. Information regarding sample size by measure within diagnostic classes available upon request. Comparisons for OCD diagnostic groups not included due to inadequate sample size ($n < 10$; Kraemer, 1992). Mean scores are rounded to the first decimal place. Differences between scale scores by diagnostic group are quantified as effect sizes (Cohen's *d*; Cohen, 1960). Formula for comparison of effect size within groups derived from Hasselbad & Hedges (1995). Cohen's d of .2 constitutes a small effect size, .5 medium effect size, and .8 large effect size for the social sciences (Cohen, 1978). Corresponding measures: *1. AAQ; 2. ASI-3; 3. BAS; 4. BIS; 5. DASS-D; 6. DASS-A; 7. DASS-S; 8. DASS-TOT; 9. DERS; 10. DIS; 11. EAT; 12. FSS-III; 13. IUS; 14. OCI-R; 15. PCL; 16. PSWQ; 17. SPS; 18. SIAS; 19. WBSI.*

Table 5
Diagnostic Efficiency Outcomes Reported for ROC Analyses

Abbreviation and Terminology		Definition
AUROC	Area Under ROC Curve	Overall measure of performance: Total area under the Receiver Operating Characteristics curve
95% C.I.	95% Confidence Interval	Values for AUROC CI at 95%
SE	Standard Error	Standard Error of the AUROC
AS	Asymptotic *p*-value	Non-parametric *p*-value statistic
Kappa	Cohen's kappa	Weighted (i.e., calibrated) measure of efficiency
Cut[1]	Ideal Cut-Point	Cut point that maximizes overall classification accuracy
Sens	Sensitivity	Proportion of participants accurately classified as positive
Spec	Specificity	Proportion of participants accurately classified as negative
Eff	Efficiency	Quantifies overall percentage of participants accurately classified (i.e., both negative and positive)
PPV	Positive Predictive Value	Probability that the disorder is present when score is at or above given cut-point
NPV	Negative Predictive Value	Probability that the disorder is not present when score is at or below given cut-point
Chi	Chi Squared	Measure of the degree of association with the outcome criterion
Sig[2]	Significance	Level of significance achieved (*p*-value)

Note. Refer back to this table for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes to follow (Table 6 through Table 16).

[1]Unless denoted otherwise ideal cut-point is greater than or equal to (≥) value provided. Italicized values represent cut-points that are < (i.e., *less than*) the corresponding value.

[2]Significance level: **p* < .01; ***p* < .001; ****p* < .0001.

References: Kraemer, 1992; Hanley & McNeil, 1982, 1983; Brown & Davis; Cohen, 1968; Fawcett, 2004; Green & Swets, 1996; Linden, 2006; McFall & Treat, 1999; Metz, 1978; Murphy et al., 1987; Pepe, 2003; Streiner & Cairney, 2007; Swets, 1988; Swets & Pickett, 1982; Tripepi, Jager, Dekker, & Zoccali, 2009; Whiting et al., 2004; Zhou, Obuchowski, & McClish, 2002.

Table 6
Diagnostic Efficiency: ≥1 Psychiatric Diagnosis

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.650	.585-.715	.033	.000	.202	22	.710	.495	.637	.73	.47	11.92	***
ASI-3		.687	.622-.751	.033	.000	.303	7	.767	.535	.688	.76	.54	26.80	***
BIS/BAS	BIS	.621	.555-.687	.034	.001	.173	15	.477	.727	.562	.77	.42	11.27	***
	BAS	.417	.348-.487	.035	.021	.153	*27*	.554	.616	.575	.74	.41	7.62	**
DASS Total		.676	.613-.738	.032	.000	.287	9	.596	.727	.640	.81	.48	27.35	***
Depression		.652	.588-.715	.032	.000	.247	2	.632	.636	.634	.77	.47	19.00	***
Anxiety		.660	.596-.725	.033	.000	.228	2	.570	.687	.610	.78	.45	17.28	***
Stress		.635	.571-.699	.033	.000	.194	6	.456	.778	.565	.80	.42	15.23	***
DERS		.632	.566-.698	.034	.000	.206	27	.845	.343	.675	.71	.53	13.51	***
DIS		.539	.470-.608	.035	.275	.072	15	.477	.606	.521	.70	.37	1.81	
EAT		.564	.497-.632	.034	.072	.139	7	.513	.646	.558	.74	.41	6.70	**
FSS		.664	.582-.747	.042	.000	.146	25	.611	.545	.589	.72	.42	6.53	*
IUS		.682	.617-.746	.033	.000	.263	14	.756	.505	.671	.75	.52	20.17	***
OCI-R		.663	.580-.746	.042	.000	.159	3	.606	.566	.592	.73	.42	7.80	**
PCL		.639	.560-.718	.040	.002	.129	12	.290	.879	.490	.82	.39	10.45	**
PSWQ		.655	.577-.732	.039	.000	.143	33	.352	.828	.514	.80	.40	10.34	**
SIAS		.735	.665-.805	.036	.000	.196	21	.415	.828	.555	.82	.42	17.39	***
SPS		.754	.683-.826	.036	.000	.221	5	.580	.667	.610	.77	.45	15.97	***
WBSI		.658	.594-.722	.033	.000	.246	31	.580	.697	.620	.79	.46	20.14	***

Note. Prevalence = 66.1%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 7
Diagnostic Efficiency: ≥ 2 Psychiatric Diagnoses

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.661	.596-.726	.033	.000	.279	26	.618	.670	.651	.53	.74	23.18	***
ASI-3		.645	.580-.710	.033	.000	.226	11	.591	.643	.623	.50	.72	15.17	***
BIS/BAS	BIS	.611	.545-.677	.034	.002	.089	15	.527	.665	.613	.49	.70	10.48	**
	BAS	.483	.414-.552	.035	.629	.180	20	.145	.923	.630	.53	.64	3.49	
DASS Total		.674	.608-.739	.033	.000	.361	11	.636	.731	.695	.59	.77	38.27	***
Depression		.631	.564-.699	.034	.000	.247	3	.545	.703	.644	.53	.72	17.83	***
Anxiety		.643	.577-.710	.034	.000	.239	3	.518	.720	.644	.53	.71	16.66	***
Stress		.647	.580-.714	.034	.000	.242	6	.527	.714	.644	.53	.71	17.04	***
DERS		.629	.563-.695	.033	.000	.231	55	.336	.874	.671	.62	.69	18.52	***
DIS		.523	.455-.592	.035	.505	.060	12	.764	.308	.479	.40	.68	1.73	
EAT-26		.564	.495-.633	.035	.068	.134	10	.382	.747	.610	.48	.67	5.42	*
FSS		.621	.541-.701	.041	.005	.102	86	.327	.769	.603	.46	.65	3.26	
IUS		.639	.573-.705	.034	.000	.219	43	.282	.912	.675	.66	.68	19.09	***
OCI-R		.620	.539-.702	.041	.006	.108	*0*	.391	.714	.592	.45	.66	3.46	
PCL		.701	.623-.779	.040	.000	.180	13	.318	.846	.647	.56	.67	1.94	***
PSWQ		.698	.619-.777	.040	.000	.192	39	.282	.890	.661	.61	.67	14.06	***
SIAS		.600	.518-.683	.042	.021	.114	*0*	.391	.720	.596	.46	.66	3.85	*
SPS		.689	.613-.766	.039	.000	.150	11	.409	.736	.613	.48	.67	6.67	**
WBSI		.641	.574-.708	.034	.000	.273	39	.409	.846	.682	.62	.70	23.82	***

Note. Prevalence = 37.7%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 8
Diagnostic Efficiency: Major Depressive Disorder (MDD)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.653	.529-.777	.063	.026	.141	33	.263	.912	.870	.17	.95	6.10	*
ASI-3		.710	.594-.827	.059	.002	.197	21	.526	.839	.818	.19	.96	15.71	***
BIS/BAS	BIS	.531	.388-.674	.073	.655	.046	18	.211	.861	.818	.10	.94	0.73	
	BAS	.380	.267-.493	.058	.080	.061	*26*	.632	.586	.589	.10	.96	3.44	
DASS Total		.764	.644-.883	.061	.000	.224	24	.368	.916	.880	.23	.95	15.56	***
Depression		.760	.639-.880	.061	.000	.241	11	.263	.960	.914	.31	.95	17.04	***
Anxiety		.689	.583-.794	.054	.006	.106	4	.526	.740	.726	.12	.96	6.28	*
Stress		.739	.626-.852	.058	.000	.215	12	.316	.930	.890	.24	.95	13.75	***
DERS		.528	.425-.631	.053	.685	.052	42	.632	.564	.568	.09	.96	2.75	
DIS		.534	.403-.665	.067	.622	.070	22	.105	.956	.901	.14	.94	1.46	
EAT-26		.510	.370-.651	.072	.880	.053	*2*	.263	.832	.795	.10	.94	1.10	
FSS		.557	.426-.688	.067	.478	.054	99	.316	.795	.764	.10	.94	1.30	
IUS		.638	.502-.773	.069	.045	.156	49	.316	.897	.860	.18	.95	7.85	**
OCI-R		.563	.400-.726	.083	.434	.093	24	.263	.875	.836	.13	.94	2.95	
PCL		.727	.601-.853	.064	.005	.203	24	.368	.905	.870	.21	.95	13.22	***
PSWQ		.686	.539-.833	.075	.020	.126	37	.474	.795	.774	.14	.96	7.40	**
SIAS		.479	.348-.609	.067	.790	.033	21	.421	.674	.658	.08	.94	0.72	
SPS		.539	.400-.678	.071	.627	.091	29	.211	.905	.860	.13	.94	2.56	
WBSI		.702	.593-.810	.055	.003	.148	38	.632	.744	.736	.15	.97	12.38	***

Note. Prevalence = 6.5%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 9
Diagnostic Efficiency: Generalized Anxiety Disorder (GAD)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.690	.615-.765	.038	.000	.243	26	.741	.630	.651	.31	.91	24.61	***
ASI-3		.645	.567-.723	.040	.001	.182	21	.333	.849	.753	.33	.85	9.68	**
BIS/BAS	BIS	.545	.547-.632	.045	.305	.130	18	.241	.878	.760	.31	.84	5.05	*
	BAS	.384	.298-.470	.044	.008	.204	*20*	.241	.929	.801	.43	.84	13.69	***
DASS Total		.800	.744-.855	.028	.000	.372	11	.852	.693	.723	.39	.95	54.17	***
Depression		.751	.680-.822	.036	.000	.237	3	.667	.672	.671	.32	.90	21.25	***
Anxiety		.735	.671-.799	.032	.000	.345	5	.537	.840	.784	.43	.89	35.45	***
Stress		.770	.703-.838	.034	.000	.344	7	.667	.769	.750	.40	.91	38.93	***
DERS		.623	.543-.702	.041	.005	.194	55	.370	.832	.747	.33	.85	11.03	***
DIS		.552	.470-.634	.042	.232	.080	19	.204	.870	.747	.26	.83	1.93	
EAT-26		.545	.463-.627	.042	.300	.079	21	.148	.916	.774	.29	.83	2.09	
FSS		.625	.527-.723	.050	.022	.108	86	.370	.756	.685	.26	.84	3.61	
IUS		.676	.595-.758	.042	.000	.246	52	.259	.941	.815	.50	.85	2.40	***
OCI-R		.612	.516-.707	.049	.041	.077	7	.500	.613	.592	.23	.84	2.35	
PCL		.644	.540-.749	.053	.008	.163	27	.204	.929	.795	.39	.84	8.88	**
PSWQ		.707	.607-.804	.050	.000	.224	43	.278	.916	.798	.43	.85	15.66	***
SIAS		.631	.531-.731	.051	.016	.111	34	.222	.878	.757	.29	.83	3.67	
SPS		.635	.532-.739	.053	.013	.133	27	.222	.895	.771	.32	.84	5.46	*
WBSI		.678	.605-.750	.037	.000	.214	37	.519	.748	.705	.32	.87	14.84	***

Note. Prevalence = 18.5%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 10
Diagnostic Efficiency: Panic Disorder (PD)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.710	.607-.814	.053	.001	.191	36	.190	.967	.911	.31	.94	11.33	***
ASI-3		.712	.603-.820	.055	.001	.201	23	.476	.852	.825	.20	.95	14.83	***
BIS/BAS	BIS	.644	.515-.773	.066	.028	.134	16	.571	.742	.729	.15	.96	9.46	**
	BAS	.407	.280-.535	.065	.157	.092	*23*	.429	.768	.743	.13	.95	4.03	*
DASS Total		.750	.668-.833	.042	.000	.165	17	.524	.801	.781	.17	.96	11.86	***
Depression		.685	.572-.797	.057	.005	.134	6	.381	.841	.808	.16	.95	6.68	**
Anxiety		.750	.645-.854	.053	.000	.243	5	.667	.812	.801	.22	.97	25.78	***
Stress		.680	.574-.786	.054	.006	.104	12	.19	.923	.870	.16	.94	3.18	
DERS		.669	.562-.776	.055	.010	.132	60	.286	.889	.846	.17	.94	5.52	*
DIS		.677	.562-.792	.059	.007	.212	21	.238	.956	.904	.29	.94	13.35	***
EAT-26		.582	.442-.772	.071	.211	.147	19	.286	.900	.856	.18	.94	6.73	**
FSS		.618	.433-.802	.094	.156	.198	143	.286	.930	.884	.24	.94	11.57	***
IUS		.695	.569-.822	.065	.003	.357	63	.286	.985	.935	.60	.95	43.26	***
OCI-R		.672	.518-.827	.079	.038	.145	39	.143	.970	.911	.27	.94	6.91	**
PCL		.671	.495-.847	.090	.040	.265	38	.238	.974	.921	.42	.94	22.28	***
PSWQ		.654	.498-.809	.079	.064	.181	53	.190	.963	.908	.29	.94	10.07	**
SIAS		.667	.526-.808	.072	.045	.145	48	.143	.970	.911	.27	.94	6.91	**
SPS		.708	.544-.872	.084	.012	.269	35	.286	.959	.911	.35	.95	21.36	***
WBSI		.730	.628-.832	.052	.000	.196	43	.476	.849	.822	.20	.95	14.27	***

Note. Prevalence = 7.2%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 11
Diagnostic Efficiency: Post-Traumatic Stress Disorder (PTSD)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.672	.539-.804	.068	.030	.118	31	.429	.838	.818	.12	.97	6.58	*
ASI-3		.680	.548-.812	.067	.023	.109	29	.214	.928	.894	.13	.96	3.72	
BIS/BAS	BIS	.641	.502-.780	.071	.076	.092	16	.571	.734	.726	.10	.97	6.15	*
	BAS	.564	.426-.702	.070	.419	.045	30	.429	.723	.709	.07	.96	1.51	
DASS Total		.710	.560-.861	.077	.008	.222	24	.429	.914	.890	.20	.97	16.94	***
Depression		.733	.584-.881	.076	.003	.214	7	.500	.888	.870	.18	.97	17.77	***
Anxiety		.609	.444-.775	.084	.167	.103	6	.429	.82	.801	.11	.97	5.32	*
Stress		.675	.510-.839	.084	.027	.208	12	.357	.928	.901	.20	.97	13.85	***
DERS		.690	.559-.822	.067	.016	.130	62	.286	.910	.880	.14	.96	5.71	*
DIS		.540	.388-.393	.078	.609	.037	17	.429	.701	.688	.07	.96	1.06	
EAT-26		.651	.530-.772	.062	.057	.073	6	.857	.532	.548	.08	.99	8.10	**
FSS		.471	.333-.610	.071	.751	.045	27	.786	.475	.490	.07	.98	3.64	
IUS		.599	.446-.752	.078	.210	.106	49	.286	.892	.863	.12	.96	4.10	*
OCI-R		.503	.084-.666	.084	.976	.052	30	.143	.924	.887	.09	.96	0.83	
PCL		.873	.775-.970	.050	.000	.363	31	.500	.950	.928	.33	.97	40.37	***
PSWQ		.598	.423-.773	.089	.276	.195	54	.214	.968	.932	.25	.96	11.19	***
SIAS		.538	.392-.685	.075	.671	.050	17	.643	.59	.592	.07	.97	2.96	
SPS		.568	.392-.744	.090	.446	.098	32	.214	.921	.887	.12	.96	3.11	
WBSI		.722	.578-.866	.073	.005	.168	42	.571	.831	.818	.15	.97	14.12	***

Note. Prevalence = 4.8%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 12
Diagnostic Efficiency: Social Anxiety Disorder (SAD)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.617	.551-.682	.033	.001	.175	31	.269	.890	.637	.63	.64	12.38	***
ASI-3		.667	.605-.730	.032	.000	.254	12	.555	.699	.640	.56	.70	18.90	***
BIS/BAS	BIS	.656	.591-.720	.033	.291	.291	15	.580	.711	.658	.58	.71	24.70	***
	BAS	.437	.370-.504	.034	.134	.104	*28*	.630	.480	.541	.45	.65	3.47	
DASS		.607	.540-.674	.034	.002	.222	9	.622	.607	.613	.52	.70	14.77	***
Depression		.586	.519-.652	.034	.013	.145	2	.571	.578	.575	.48	.66	6.31	*
Anxiety		.601	.534-.668	.034	.003	.193	3	.504	.688	.613	.53	.67	1.93	***
Stress		.589	.522-.657	.035	.009	.175	6	.479	.694	.606	.52	.66	8.95	**
DERS		.609	.544-.674	.033	.002	.179	56	.286	.879	.637	.62	.64	12.45	***
DIS		.490	.422-.558	.035	.770	.045	20	.126	.913	.592	.50	.60	1.18	
EAT-26		.526	.457-.595	.035	.453	.117	7	.529	.590	.565	.47	.65	4.02	*
FSS		.697	.623-.771	.038	.000	.222	73	.445	.769	.637	.57	.67	14.90	***
IUS		.666	.603-.728	.032	.000	.244	21	.639	.613	.623	.53	.71	17.83	***
OCI-R		.627	.548-.705	.040	.003	.146	11	.395	.746	.603	.52	.64	6.50	*
PCL		.627	.546-.708	.041	.003	.127	13	.286	.832	.61	.54	.63	5.81	*
PSWQ		.650	.570-.730	.041	.000	.169	38	.294	.861	.63	.59	.64	1.56	**
SIAS		.716	.642-.790	.038	.000	.245	26	.361	.867	.661	.65	.66	21.02	***
SPS		.761	.695-.827	.034	.000	.251	11	.462	.780	.651	.59	.68	19.11	***
WBSI		.596	.529-.662	.034	.005	.183	35	.462	.717	.613	.53	.66	9.85	**

Note. Prevalence = 40.7%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 13
Diagnostic Efficiency: Specific Phobia (SP)

Index Test		AUROC	95% C.I.	SE	AS	Kappa .5	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.592	.525-.658	.034	.008	.159	25	.579	.585	.582	.50	.66	7.57	**
ASI-3		.623	.558-.688	.033	.000	.232	10	.636	.602	.616	.53	.70	16.15	***
BIS/BAS	BIS	.561	.495-.628	.034	.074	.115	14	.587	.532	.555	.47	.65	4.01	*
	BAS	.533	.466-.600	.034	.341	.118	24	.785	.345	.527	.46	.69	5.82	*
DASS Total		.602	.536-.668	.034	.003	.174	18	.289	.871	.630	.61	.63	11.63	***
Depression		.553	.486-.620	.034	.122	.094	5	.281	.807	.589	.51	.61	3.10	
Anxiety		.605	.538-.672	.034	.002	.228	5	.347	.865	.651	.65	.65	18.51	***
Stress		.583	.516-.650	.034	.015	.168	9	.264	.889	.630	.63	.63	11.56	***
DERS		.590	.524-.657	.034	.008	.170	55	.298	.86	.627	.60	.63	10.72	**
DIS		.532	.464-.600	.035	.347	.105	17	.364	.737	.582	.49	.62	3.38	
EAT-26		.570	.502-.638	.035	.041	.155	9	.438	.713	.599	.52	.64	7.15	**
FSS		.617	.536-.697	.041	.006	.150	98	.298	.842	.616	.57	.63	8.16	**
IUS		.604	.539-.670	.033	.002	.165	27	.479	.684	.599	.52	.65	8.02	**
OCI-R		.644	.566-.723	.040	.001	.150	14	.322	.819	.613	.56	.63	7.73	**
PCL		.646	.564-.727	.042	.001	.178	12	.331	.836	.627	.59	.64	11.04	***
PSWQ		.618	.537-.699	.041	.005	.137	39	.248	.877	.616	.59	.62	7.70	**
SIAS		.594	.515-.674	.041	.025	.091	23	.339	.749	.579	.49	.62	2.64	
SPS		.625	.546-.705	.041	.003	.122	32	.149	.959	.623	.72	.61	10.52	**
WBSI		.606	.539-.673	.034	.002	.208	38	.397	.801	.634	.59	.65	13.74	***

Note. Prevalence = 41.4%
Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 14
Diagnostic Efficiency: Eating Disorders (ED)

Index Test		AUROC	95% C.I.	SE	AS	Kappa	Cut	Sens	Spec	Eff	PPV	NPV	Chi	Sig
AAQ		.473	.370-.576	.052	.607	.081	*17*	.200	.883	.801	.19	.89	1.93	
ASI-3		.515	.402-.627	.057	.779	.148	25	.286	.875	.805	.24	.90	6.50	*
BIS/BAS	BIS	.525	.431-.619	.048	.628	.039	14	.600	.494	.507	.14	.90	1.09	
	BAS	.474	.374-.575	.051	.622	.078	*17*	.086	.969	.863	.27	.89	2.53	
DASS Total		.559	.453-.666	.054	.255	.106	18	.314	.821	.760	.19	.90	3.59	
Depression		.548	.447-.648	.051	.362	.064	10	.114	.938	.839	.20	.89	1.31	
Anxiety		.585	.479-.692	.054	.101	.171	5	.429	.805	.760	.23	.91	9.75	**
Stress		.517	.415-.618	.052	.746	.05	8	.314	.759	.705	.15	.89	0.88	
DERS		.560	.458-.661	.052	.252	.092	54	.343	.786	.733	.18	.90	2.89	
DIS		.535	.433-.638	.052	.499	.049	15	.543	.564	.562	.15	.90	1.43	
EAT-26		.569	.469-.669	.051	.186	.170	27	.143	.977	.877	.45	.89	12.14	***
FSS		.543	.379-.706	.084	.634	.231	*0*	.686	.720	.716	.25	.94	22.96	***
IUS		.523	.422-.624	.052	.658	.080	53	.143	.926	.832	.21	.89	1.94	
OCI-R		.580	.440-.719	.071	.375	.235	*0*	.686	.724	.719	.25	.94	23.53	***
PCL		.636	.480-.792	.079	.130	.224	*0*	.686	.712	.709	.24	.94	21.86	***
PSWQ		.567	.416-.717	.077	.458	.239	*0*	.686	.728	.723	.26	.94	24.11	***
SIAS		.570	.396-.744	.089	.437	.239	*0*	.686	.728	.723	.26	.94	24.11	***
SPS		.679	.532-.825	.075	.047	.235	*0*	.686	.724	.719	.25	.94	23.53	***
WBSI		.580	.477-.683	.052	.123	.107	37	.457	.720	.688	.18	.91	4.58	*

Note. Prevalence = 12.0%; Anorexia nervosa (n=4), bulimia nervosa (n=14), and eating disorder not otherwise specified (n=17) were collapsed into a single “eating disorder” category (n=35). Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.

Table 15
Best Performing Measure by Diagnostic Category

	n	P(%)	Best Measure	Cut	AUROC	Kappa	Sens	Spec	Eff	PPV	NPV	Chi*
≥ 1 Disorder	193	66.1	ASI-3	7	.687	.303	.767	.535	.688	.76	.54	26.80
≥ 2 Disorders	110	37.7	DASS-Tot	11	.674	.361	.636	.731	.695	.59	.77	38.27
GAD	54	18.4	DASS-Tot	11	.800	.372	.852	.639	.723	.39	.95	54.17
MDD	19	6.5	DASS-Dep	11	.760	.241	.263	.960	.914	.31	.95	17.04
PD	21	7.2	IUS	63	.695	.357	.286	.985	.935	.60	.95	43.26
PTSD	14	4.8	PCL	31	.873	.363	.500	.950	.928	.33	.97	40.37
SAD	119	40.7	BIS	15	.656	.291	.580	.711	.658	.58	.71	24.70
SP	121	41.4	ASI	10	.623	.232	.636	.602	.616	.53	.70	16.15
ED	35	12.0	EAT-26	27	.569	.170	.143	.977	.877	.45	.89	12.14

Note. *n* = number of true positive cases as identified by gold standard. P = prevalence in current sample. Refer to Table 5 for definitions and explanation of outcomes and notations contained within tables reported diagnostic efficiency outcomes.
*All Chi-squared values for index test cut-offs reported in this table were significant at the ***$p < .0001$ level.

VII. CURRICULUM VITA

Lindsay R. Trent, M.A.

EDUCATION

B.A. in Psychology *(May 2008)*
University of Tennessee at Knoxville

M.A. Clinical Psychology *(August 2010)*
University of Mississippi
Advisor: John Young, Ph.D.
Thesis: *Development of a Measure of Disseminability (MOD)*
Committee: Alan Gross, Ph.D. (Chair), John Young, Ph.D., & Erin Buchanan, Ph.D.

Doctoral Candidate Clinical Psychology (*Present)*
University of Mississippi
Advisor: John Young, Ph.D.
Degree Conferment Anticipated: August 2014
Dissertation: *Diagnostic Accuracy of Self-Report Instruments in a Nonclinical Sample: A Receiver Operating Characteristics (ROC) analysis*
Committee: John Young, Ph.D., Alan Gross, Ph.D., Danielle Maack, Ph.D., & John Bentley, Ph.D.

Predoctoral Intern Clinical Psychology *(August 2013-Present)*
Birmingham VA/University of Alabama Birmingham (UAB)
Supervisors: Hal Thurstin, Ph.D. & Chebon Porter, Ph.D.
Expected Completion: July 31st, 2014

RESEARCH EXPERIENCE

Mississippi Youth Around the Clock (MYPAC)
Co-Project Grant Coordinator
University of Mississippi/Mississippi Children's Home Society
PI/Supervisor: John Young, Ph.D.
August 2008 - August 2009

- Clinical and Program Evaluation Services for Mississippi Children's Home Society
- Co-organized and implemented a grant for child mental and behavioral health assessments
- Site recruitment (i.e., outreach efforts), scheduling, administration, data management and interpretation

Understanding Clinical Supervision
Co-Investigator
University of Mississippi
PI/Supervisor: John Young, Ph.D.
August 2008 - May 2009

- Conducted content analysis of videotaped supervision meetings to inform iterative development of a supervision coding instrument
- Supervised and trained undergraduate research assistants to ensure codification fidelity

Development of a Measure of Disseminability (MOD)
Principal Investigator
University of Mississippi
Supervisor: John Young, Ph.D.
October 2008 - December 2008

- Developed a measure assessing dissemination-relevant variables following published guidelines for iterative development and psychometric validation
- Conducted psychometric examination employing exploratory factor analytic methods

Dissemination and Implementation Coding in Child and Adolescent Randomized Controlled Trial (RCT) Literature
Coding Consultant
Practicewise, LLC
Supervisors: Bruce Chorpita, Ph.D., Chad Ebesutani, Ph.D., & Eric Daleiden, Ph.D.
October 2008 - Present

- Codifying outcome data across randomized controlled trials informing a modularized treatment approach
- Independently coded over 100 child-focused RCTs applying a detailed, iterative coding scheme to extant RCTs in relevant literature

Psychometric replication of a Measure of Disseminability (MOD) in an independent sample
Principal Investigator
University of Mississippi
Supervisor: John Young, Ph.D.
September 2010 - December 2010

- Examined the reproducibility of the Measure of Disseminability's (MOD) psychometric properties in an independent sample applying confirmatory factor analytic methods

Attitudes about treatment and mental health service selection (Stakeholder Liaison Research Panel)
Co-Investigator
University of Mississippi
PI/Supervisor: John Young, Ph.D.
May 2012 - Present

- Part of a multi-site study assessing therapy-related attitudes of clients within frontline care settings
- Quantitative and qualitative data collection driving iterative development of an empirically-derived construct definition and assessment strategy

Diagnostic accuracy of self-report instruments in a non-clinical sample of undergraduates: A Receiver Operating Characteristics (ROC) analysis
Principal Investigator
University of Mississippi
Supervisor: John Young, Ph.D.
August 2012 - Present

- Used ROC analyses to examine the diagnostic validity of freely-available assessment tools
- Reported indices of diagnostic validity including optimal cut points for diagnostic screening purposes

NIH Sponsored Smoking Cessation Research Trial
Research Trainee
University of Alabama (UAB)
Supervisor: Karen Cropsey, Psy.D.
August 2013 – Present

- NIH Sponsored trial using Bupropion SR plus two different types of counseling for smoking cessation
- Manuscript preparation, data analysis, and grant-writing activities

SAMHSA Sponsored Behavioral Day Treatment Program
Research Trainee
University of Alabama (UAB)
Supervisors: Jesse Milby, Ph.D. & Joseph Schumacher, Ph.D.
August 2013 – Present

- SAMHSA sponsored translational research project treating co-occurring substance use and SMI in homeless a population utilizing Therapeutic Goal Management (TGM) and Contingency Managed Housing manualized interventions with demonstrated efficacy
- Implementation activities in active research trial (i.e., therapy provision, participant screening, developed revised version of treatment and fidelity protocol)
- Data analysis and manuscript preparation from the *Homeless Four* database

PROFESSIONAL PUBLICATIONS

1. Chorpita, B.F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., Phillips, L., Hershberger, A., Stumf, R., **Trent, L.,** Smith, R. L., Okamura, K., & Starce, N. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice, 18,* 153-171.

2. Carvahlo, J., & **Trent, L. R. (**2011). The impact of decreased environmental reward in predicting depression severity: Support for behavioral theories of depression. *Psychopathology, 44*(4), 242-252.

3. Viana, A. G., **Trent, L**., Tull, M. T., Heiden, L., Damon, J. D., Hight, T. L., & Young, J. (2012). Non-medical use of prescription drugs among Mississippi youth: Constitutional, psychological, and family factors. *Addictive Behaviors, 37,* 1382 - 1388.

4. Trent, L. R., Buchanan, E., Ebesutani, C., Ale, C., Heiden, L., Hight, T. et al. (2013). A measurement invariance examination of the Revised Child Anxiety and Depression Scale in a southern sample: Differential item functioning between African American and Caucasian youth. *Assessment, 20,* 175-187.

5. Cropsey, K. L., Levanthal, A., Stevens, E., **Trent, L. R.,** Clark, B., Hardy, S., et al. (2014). Expectancies for the effectiveness of different tobacco interventions account for racial and gender differences in motivation to quit and abstinence self-efficacy. *Nicotine & Tobacco Research.* Advance online publication. doi:10.1093/ntr/ntu048

6. Cropsey. K. L., **Trent, L. R**., Stevens, E., & Clark, B. (2014). How low should you go? Determining the optimal cut-off for exhaled carbon monoxide using cotinine as reference. *Nicotine & Tobacco Research.* Advance online publication. doi: 10.1093/ntr/ntu085

7. Ebesutani, C., Fierstein, M. Viana, A. G., **Trent, L.,** & Young, J. (In press). The mediating role of loneliness in the relationship between anxiety and depression in youth. *Psychology in the Schools*.

MANUSCRIPTS SUBMITTED/UNDER REVIEW

8. Chin, E., Drescher, C. F., **Trent, L. R**., Darden, M. C., Vosbein, M., & Johnson, L. R. A direct comparison: Factor structure of the English and Chinese language versions of the 12-item General Health Questionnaire (GHQ-12) in Malaysia.

9. Stevens, E., Cropsey, K. L., **Trent, L. R.,** Clark, B., & Lahti, A. H. Gender differences in the associations between neurocognitive abilities and depression for smokers in the criminal justice system.

10. Cropsey. K. L., Nair, P., Clark, B., **Trent, L. R**., Stevens, E., & Binswanger, I. Factors associated with opioid relapse following release from jail.

11. Clark, B., Hendricks, P., Lane, P., Cropsey, K.L., & **Trent, L. R**. Methadone maintenance treatment may improve completion rates and delay opioid relapse for opioid dependent individuals under community correction supervision.

MANUSCRIPTS IN PREPARATION

12. Trent, L. R., Freeman, A., Ebesutani, C., & Young, J. Diagnostic Accuracy of Self- Report Instruments in a Community Sample: A Receiver Operating Characteristics (ROC) analysis.

13. Trent, L. R., Buchanan, E., Ebesutani, C., & Young, J. Development and psychometric replication of the Measure of Disseminability (MOD).

14. Trent, L. R., Cropsey, K. L., Stevens, E., & Clark, B. Using baseline characteristics to predict differential treatment response in a community-focused smoking cessation trial.

PROFESSIONAL TALKS & PRESENTATIONS

1. Trent, L., & Hindman, L. (2010, November). *Assessing Mental Health Outcomes in Mississippi Youth: The Behavioral Vital Signs Project.* Invited talk given at the Louisiana Counseling Association Conference, Baton Rouge, LA.

2. Trent, L. (2011, November). *Development of a Measure of Disseminability.* Talk given at Association of Behavioral and Cognitive Therapy, Toronto, Canada.

3. Trent, L., & Young, J. (2009, November). *Clinical supervision: A field-wide survey.* Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, New York, NY.

4. Trent, L., & Young, J. (2009, November). *Development of a measure of disseminability.* Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, New York, NY.

5. Flegle, L., **Trent, L.**, Ambrose, C., Latzman, R.L., & Young, J. (2009, November). *Predictors of clinical elevations from a school-based mental health screening.* Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, New York, NY.

6. Trent, L., Drescher, C., & Young, J. (2010, March). *How do we know what sells? Developing a Measure of Dissemination for Mental Health Treatments.* Poster presentation at the 3rd annual NIH Conference on the Science of Dissemination and Implementation, Bethesda, MY.

7. Chin, E., Drescher, C., **Trent, L.,** Ambrose, A., Heiden, L., & Young, J. (2010, October). *Dissemination in school systems: Feedback from Behavioral Vital Signs Personnel.* Poster presented at the Center for School Mental Health, Albuquerque, NM.

8. Trent, L., Buchanan, E., & Young, J. (2010, November). Development and initial psychometric examination of the Measure of Disseminability. Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, San Francisco, CA (Dissemination and Implementation Special Interest Group).

9. Trent, L., Drescher, C., Buchanan, E., & Young, J. (2010, November). The Measure of Disseminability: Confirmatory factor analysis and psychometric replication. Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, San Francisco, CA (Dissemination and Implementation Special Interest Group).

10. Drescher, C. F., **Trent, L.** R., Heiden, L., Hight, T., Damon, J. D., & Young, J. (2010, November). Factors Affecting Dissemination of a Youth Mental Health Screening. Poster presentation at the annual meeting of the Association for Behavioral and Cognitive Therapies, San Francisco, CA (Dissemination and Implementation Special Interest Group).

11. Chin, E., **Trent, L.,** & Young, J. (2012, November). *Examining the 21-item Depression, Anxiety, and Stress Scales with Receiver Operating Characteristic curves in a college sample.* Poster will be presented at the 2012 Association for Behavioral and Cognitive Therapies, National Harbor - MD.

12. Drescher, C. F., Chin, E., **Trent, L. R**., Darden, M. C., Vosbein, M., & Johnson, L. R. (2013, November). *An Analysis of the Psychometric Properties of the English and Chinese Versions of the Meaning in Life Questionnaire (MLQ): A Malaysian College Sample.* Poster to be presented at the annual meeting of the Association for Behavioral and Cognitive Therapies, Nashville, TN.

13. Drescher, C. F., Chin, E., **Trent, L. R.,** Darden, M. C., Vosbein, M., Khor, K. L., Seak, R., Loo, A., Romeo, S. & Johnson, L. R. (2013, November). *An analysis of the psychometric properties of the English and Chinese versions of the Meaning in Life Questionnaire (MLQ): A Malaysian college sample.* Poster to be presented at the annual meeting of the Association for Behavioral and Cognitive Therapies, Nashville, TN.

ORGANIZATIONAL CONSULTING

Behavioral Vital Signs (BVS) Grant Project Consultant
Supervisor: John Young, Ph.D. & Laurie Heiden, M.Ed.
January 2009 – May 2013

- Implemented interventions appropriate to the specific problem areas and goals of the organization (e.g., mental health related didactic presentations)
- Data analysis and management
- Generated reports detailing outcome data for each school

TEACHING/TRAINING EXPERIENCE

Therapist Trainer
Mississippi Child Home Services (Jackson, MS)
Supervisor: John Young, Ph.D.
January 2010 – May 2010

Provided didactic training to therapists as part of the *Mississippi Youth Around the Clock (MYPAC)* Project

- Motivational Interviewing (MI) techniques (Miller & Rollnick, 2002)
- CBT-based parent training modules (MATCH-ADTC; Chorpita & Weisz, 2009)

Graduate Instructor
University of Mississippi
Supervisors: Todd Smitherman, Ph.D. & John Young, Ph.D.
July 2010 – May 2013

Primary Instructor for 12 courses

- Introduction to Psychology 101
  (6 sections)
- Learning and Behavior 309
  (2 sections)
- Abnormal Psychology 311
  (4 sections)

ACCOMPLISHMENTS & AWARDS

Outstanding Student Poster Award

Dissemination and Implementation Special Interest Group: ABCT Convention 2010 (**Trent, L.,** Buchanan, E., & Young, J., 2010). Development and initial psychometric examination of the Measure of Disseminability.

Passed Exam for Professional Practice of Psychology (EPPP) at Doctoral Level

Kentucky Board of Psychology
April, 2012

CLINICAL EXPERIENCE

Behavioral Consultant, DeSoto County Schools
Desoto, MS
Supervisor: Sheila Williamson, Ph.D.
July 2009 - July 2010

- Conducted functional behavior assessments including behavioral observation across multiple contexts, teacher and parent interviews, and provided report detailing observational findings and suggested behavioral plan
- Provided support and education for teachers and parents
- Provided individual therapy and support services for children with a range of educational disabilities and psychological disorders

Graduate Student Therapist, Psychological Services Center
University of Mississippi
Supervisors: Todd Smitherman, Ph.D., John Young, Ph.D., & Scott Gustafson, Ph.D.
August 2009 – May 2013

- Conducted psychodiagnostic evaluations
- Generated intake reports
- Provided individual and group psychotherapy for children and adults across a broad range of diagnoses
- Developed and Co-led Social Phobia Group *(December 2010 - February 2011)*

Mental Health Therapist, Communicare Community Mental Health
Oxford, MS
Supervisors: Elizabeth Dillon, Ph.D. & Dixie Church, M.A.
July 2010 - July 2011

- Conducted therapy, triage, emergency, and intake procedures with diverse clients including (e.g., low-income families, patients with disabilities, serious mental illness, and individuals enrolled in alcohol and drug treatment)
- Generated written intake reports including comprehensive treatment plans
- Consulted with members of a multidisciplinary treatment team (i.e., case managers, psychiatrists, and primary physicians)
- Conducted TB/HIV assessment and education, hospital consultation, and pre-evaluation screenings for Mississippi involuntary commitment procedures
- Served as primary therapist for on-call emergency duty during non-business hours

Neuropsychological Examiner, Le Bonheur Children's Hospital
Memphis, TN
Supervisor: Vicki Brewer, Ph.D.
February 2011 - March 2012

- Conducted comprehensive pediatric neuropsychological assessments for outpatient and inpatient hospital populations
- Diagnoses addressed included: autism spectrum disorders, epilepsy, traumatic brain injury, ADHD, neurologically based learning disabilities, and encephalopathy
- Produced integrated reports based on neurological assessments including DSM-IV-TR diagnoses and recommendations

Clinical Psychology Intern, UAB Clinical Psychological Consortium
University of Alabama Birmingham and Birmingham VA Hospital
Internship Directors: Hal Thurstin, Ph.D. & Chebon Porter, Ph.D.

Geropsychology/Behavioral Medicine Rotation, Center for Psychiatric Medicine at UAB, Birmingham, AL
Supervisors: Hal Thurstin, Ph.D. & Michelle Benjamin, PhD.
August 2013 – December 2013

- Provided comprehensive neuropsychological evaluation for inpatient and outpatient populations addressing diverse cognitive, psychological, and health-related referral questions
- Conducted organ transplant evaluations (e.g., lung, heart, liver, etc.) with patients and family members assessing readiness for transplantation assessing neuropsychological functioning, coping-skills, availability of necessary post-operative support, etc.
- Developed competency in differential diagnosis of cognitive disorders in aging populations (i.e., 65 and older)
- Conduct ongoing psychotherapy with geriatric outpatients

Substance Abuse Treatment in Homeless Individuals, Center for Preventative Medicine at UAB, Birmingham, AL
Supervisors: Jesse Milby, Ph.D. & Joseph Schumacher, Ph.D.
August 2013 – December 2013

- Individual and group therapy within the context of a grant-funded translational research project
- Developed and implemented a group therapy protocol utilizing a modularized approach integrating components of Motivational Interviewing, CBT, and DBT
- Assisted in coordination of care through inter-agency collaboration (i.e., housing provision, case managers, etc.)

PTSD Clinic (PCT), Veteran's Affairs Hospital, Birmingham, AL
Supervisors: Misti Norton, Ph.D. & Susan Rathmell, Ph.D.
August 2013 – Present

- Individual therapy with patients diagnosed with PTSD utilizing Cognitive Processing Therapy (CPT) and Prolonged Exposure (PE) manualized protocols
- Delivery of therapy using telemental health technology

Southeastern Blind Rehab Center, Veteran's Affairs Hospital, Birmingham, AL
Supervisor: Chebon Porter, Ph.D.
December 2013 – Present

- Provide initial mental health intake and individual psychosocial intervention utilizing evidence-based protocols (e.g., Cognitive Processing Therapy) within a residential treatment setting for visually impaired veterans
- Provide consultation services for members of a diverse interdisciplinary staff (e.g., nurses, optometry residents, rehabilitation specialists, etc.)

Home-Based Primary Care, Veteran's Affairs Hospital, Birmingham, AL
Supervisor: Mark Phillips, Ph.D.
December 2013 – Present

- In-home provision of psychosocial services for patients incapable of accessing traditional healthcare services due to medical-related travel restrictions
- Conduct psychodiagnostic assessment, report writing, and psychosocial interventions at veterans' homes
- Consultation and care provision within an interdisciplinary team of healthcare professionals at Community Based Outpatient Clinics (CBOCs) throughout the district served by UAB

Primary Care Mental Health Team, Veteran's Affairs Hospital, Birmingham, AL
Supervisor: Carin Eubanks, Ph.D.
December 2013 – Present

- Member of an interdisciplinary team who serves as initial mental health contact for veterans and their family members immediately following referral from primary care provider (i.e., same visit)
- Provide psychodiagnostic assessment, generate intake-reports, and refer to appropriate care providers
- Attend administrative PACT meetings targeting quality-enhancement of VA care-connection services
- Conduct short-term CBT interventions and co-facilitate group CBT protocol for Chronic Pain

MEMBERSHIP IN PROFESSIONAL ASSOCIATIONS

Association for Behavioral and Cognitive Therapies (ABCT)
(*2009-present*)
Dissemination and Implementation Sciences Special Interest Group (ABCT-SIG)
(2009-present)
American Psychological Association (APA)
(*2009-present*)

SPECIALIZED TRAINING/CLINICAL CERTIFICATIONS

Therapeutic Goal Management (TGM) Certification
NREPP Certified Substance Abuse Treatment
Supervisor: Joseph Schumacher, Ph.D.
Completed November 2013

Telemental Health (TMH) Provider
VA Certified Provider: Suicide Prevention/Emergency Care Emphasis
Supervisors: Misti Norton, Ph.D. & Chebon Porter, Ph.D.
Completed February 2014

Enhancing Caregivers' Health Through REACH VA

VA Certified Provider: Healthcare promotion intervention for caretakers of veterans with dementia or other chronic illnesses that cause behavioral difficulties
Supervisor: Chebon Porter, Ph.D.
Completed February 2014

Cognitive Processing Therapy (CPT)

VA Certified CPT Provider
Supervisors: Misti Norton, Ph.D., Susan Rathmell, Ph.D, & Fran Burnette, LCSW
Completed June 2014

PROFESSIONAL ACTIVITIES

Dissemination and Implementation Sciences: Special Interest Group (SIG)

Early Career Award Committee
Stakeholder-Liaison Research Subcommittee

Guest Journal Reviewer

Journal of Abnormal Child Psychology
Behavior Therapy
Journal of Psychopathology and Behavioral Assessment
Journal of Pain and Symptom Management

PROFESSIONAL REFERENCES

John Young, Ph.D. Assistant Professor of Psychology
Department of Psychology
The University of Mississippi
Phone: (662) 915-1775
Email: jnyoung1@olemiss.edu

Alan M. Gross, Ph.D.
Director of Clinical Training
The University of Mississippi
Phone: (662) 915-5186
E-mail: pygross@olemiss.edu

A. Hal Thurstin, Ph.D., Professor
Chief Psychologist/Director of Internship Training
Department of Psychiatry and Behavioral Neurobiology
Birmingham, AL 35233
(205) 996-7008
Email: hadrian@uab.edu