VALIDATION OF THE SPANISH SIRS: BEYOND LINGUISTIC EQUIVALENCE IN THE ASSESSMENT OF MALINGERING AMONG SPANISH SPEAKING CLINICAL POPULATIONS

Amor Alicia Correa, B.A.

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

August 2010

APPROVED:

Richard Rogers, Major Professor
Kenneth W. Sewell, Committee Member
Randall J. Cox, Committee Member
Vicki L. Campbell, Chair of the Department of Psychology
James D. Meernik, Acting Dean of the Robert B. Toulouse School of Graduate Studies

Correa, Amor Alicia. Validation of the Spanish SIRS: Beyond Linguistic Equivalence in the Assessment of Malingering among Spanish Speaking Clinical Populations. Master of Science (Psychology), August 2010, 110 pp., 20 tables, references, 155 titles.

Malingering is the deliberate production of feigned symptoms by a person seeking external gain such as: financial compensation, exemption from duty, or leniency from the criminal justice system. The Test Translation and Adaptation Guidelines developed by the International Test Commission (ITC) specify that only tests which have been formally translated into another language and validated should be available for use in clinical practice. Thus, the current study evaluated the psychometric properties of a Spanish translation of the Structured Interview of Reported Symptoms. Using a simulation design with 80 Spanish-speaking Hispanic American outpatients, the Spanish SIRS was produced reliable results with small standard errors of measurement (SEM). Regarding discriminant validity, very large effect sizes (mean Cohen's *d* = 2.00) were observed between feigners and honest responders for the SIRS primary scales. Research limitations and directions for future research are also discussed.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

Response Styles

Test-taking attitudes and particular response styles can affect the validity of test data obtained in a psychological evaluation with the potential for biasing assessment results (Rogers, 1984; Rogers, 1997; Rogers, Bagby, & Dickens, 1992). This biasing is especially true if the client responds in a deceptive manner, choosing to overreport (exaggerate) or underreport (downplay) genuine symptoms of psychological distress. Mental health professionals need to take response styles into account and incorporate methods for their detection in their psychological assessments, lest they make incorrect conclusions regarding their clients. Since the inception of standardized assessment measures that rely on a patient's self-report, early researchers agreed that assessing a client's honesty and forthrightness can be a vital part of a proper evaluation. To minimize misdiagnosis, mental health professionals should always make an attempt to determine truthfulness of responses rather than assume all questions are answered in a candid manner (Hathaway & McKinley, 1940). In fact, many standardized and widely used assessment measures, such as the Minnesota Multiphasic Personality Inventory – 2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the Personality Assessment Inventory (PAI; Morey, 2007) contain validity scales to gauge these and other response styles in an effort to determine whether an examinee's reports on a psychological measure should be trusted as accurate.

Throughout the history of psychological assessment, many different response styles have been thought to influence results. Paulhus (1984) found strong empirical support for a two-component model of socially desirable responding. The two facets of socially desirable

responses discussed by Paulhus are composed of self-deception, where individuals believe their own false reports, and impression management, when individuals consciously provide spurious responses that will make them appear favorable to others. These core facets have been studied by various researchers, albeit under different names. Whether referred to as "self-deception" and "other-deception" (Sackeim & Gur, 1979), "desirability" and "defensiveness" (Kusyszyn & Jackson, 1968), or using Paulhus' terms, the implication is that information gleaned from a self-report stands at the mercy of patients' versions of their clinical conditions.

Symptom minimization could be done unintentionally by the patient, as in *self*-deception or social desirability; it can also be purposeful, as in impression management and other-deception (Kusyszyn & Jackson, 1968; Paulhus, 1984; Paulhus, Bruce, & Trapnell, 1995; Sackeim & Gur, 1979; Whyte, Fox, & Coxell, 2006). Patients' reports of their psychological state might be either an unconscious distortion of their true symptomatology or a deliberate misrepresentation. This dichotomy parallels the non-intentional feigning of somatization disorders and the deliberate fabrication of symptoms found in factitious disorders and malingering (*Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision, DSM-IV-TR*; APA, 2000). For both of these, the chief distinguishing factor seems to be whether the client is purposely reporting false symptoms.

Rogers (1984) expanded the conceptualization of response styles to encompass four basic styles: reliable, irrelevant, defensive, and malingering. According to Rogers (1984), reliable response styles yield no negative effect on the credibility of self report data. Individuals with this approach to a psychological evaluation generally attempt to answer assessment questions honestly. For this reason, this response style is the most likely to produce valid and useful assessment results.

Most other response styles might be seen as having adverse effects on the validity of test data. For instance, an irrelevant response style yields problematic results for more than one reason. This pattern is evident when individuals appear haphazard and inconsistent with their responses to test items. It is likely that these persons are not fully engaged in the assessment process (Heilbrun, 1990; Rogers, 1997). A number of different reasons could account for this, so it is important to judge why a particular person might respond in this manner. For example, it could be that the individual is purposely not paying attention, cannot understand test items, or it could be that their symptoms are so severe that the person cannot mentally or physically engage in the assessment (Rogers, Bagby, & Dickens, 1992). Therefore clinicians must not only make a determination that assessment results are invalid, but must also attempt to determine why a particular client displays this response pattern, because the reasons might have an important effect on treatment considerations.

Defensiveness and malingering share elements of dissimulation motivated by external goals. According to Rogers, defensive individuals deny or minimize symptoms of psychological impairment. Underreporting of symptoms is a chief clinical concern because individuals engaging in this response style appear less impaired than they actually are, and could consequently be denied necessary psychological intervention (Meehl & Hathaway, 1946; Rogers & Schuman, 2005). Individuals purposely reporting false symptoms are generally thought to fall into two main diagnostic groups, specifically factitious disorders and malingering (Overholser, 1990). Individuals diagnosed with factitious disorders, fabricate symptoms for no external reward. The reward may not be obvious to persons other than the patient because incentive to feign is an internal drive, producing personal and intangible benefits (Gorman, 1982; Hagglund, 2009). The *DSM-IV-TR* narrows this conceptualization by specifying that the person's

motivation for symptom fabrication must be to assume "the sick role" and garner the attention that comes with being considered a patient (APA, 2000). On the other hand, malingering individuals purposely falsify or exaggerate symptoms for an external objective (Rogers, 2008). A more detailed analysis of the malingering response style and its importance for psychological assessment are explored in the following section.

Being suspected of intentionally reporting false symptoms, or *feigning* can have significant consequences for an examinee. In a clinical setting, this conclusion can preclude an individual from receiving mental health interventions (Rogers, 1997; Rogers, 2008) because in settings where resources are scarce, many mental health professionals believe it is their responsibility to ensure that only the truly sick receive access mental health treatment (Resnick, 1984). When encountered in a forensic setting, the ramifications can be even more serious. Not only might individuals be denied mental health care, the classification of malingering could be used to discredit them at all stages of the trial process (Rogers & Schuman, 2005). The criminal justice system attempts to ensure that only ill individuals, not malingerers, are excused from punishment. Thus, the misuse of a powerful clinical construct such as malingering can be quite damaging. Often, once individuals have been classified as malingerers, it is difficult for them to prove the genuineness of their disorders in future situations. For these reasons, a thorough assessment should be conducted before making such a classification (Berry, Baer, Rinaldo, & Wetter, 2002). Malingering and assessment methods are subsequently discussed in further detail.

Malingering

As mentioned earlier, *DSM-IV-TR* (APA, 2000) identifies malingering as the deliberate production of feigned symptoms by a person seeking some form of external gain. Other *DSM*

diagnoses, such as factitious disorders and somatoform disorders, also involve the production of false symptoms, but the key difference is the underlying motive. According to APA, (2000) only malingerers intentionally falsify symptoms for the purpose of obtaining an obvious external benefit, such as financial compensation, exemption from duty, or leniency from the criminal justice system. Possible motives for falsifying symptoms are wide and varied; therefore, malingering is encountered in a variety of clinical and forensic situations (Reid, 2000). Malingering can be difficult to detect accurately because an individual's method of feigning can differ extensively from client to client.

Criticisms of the *DSM-IV-TR* definition and disagreement among researchers further complicate professionals' ability to accurately classify malingering (DeClue, 2002). Disagreements in the field can lead to confusion regarding important points of focus during a comprehensive assessment, as well as obfuscate what information is necessary for accurate classification. For instance, the broad *DSM* definition stated above is generally accepted, but experts in malingering do not always agree that the specific criterion points made by the APA's diagnostic manual are sufficient. The criteria outlined by the *DSM-IV* is as follows (APA, 2000, p.739):

> "Malingering should be strongly suspected if any combination of the following is noted:
>
> 1. Medico-legal context of presentation (e.g., the person is referred by an attorney to the clinician for examination)
> 2. Marked discrepancy between the person's claimed stress or disability and the objective findings
> 3. Lack of cooperation during the diagnostic evaluation and in complying with the prescribed treatment regimen
> 4. The presence of Antisocial Personality Disorder

Some professionals advocate that the *DSM-IV* indices provide good guidelines for identifying potential malingerers during an assessment (Meyer & Deitsch, 1996). Others, such as

Rogers and Schuman (2005), contest this viewpoint. They note that most patients undergoing a forensic evaluation will meet several of these indices even if they are not malingering, simply due to the nature of the assessment. Specifically, all criminal defendants will meet the first criterion point (i.e., medico-legal context). It is likely that a majority of criminal defendants will also qualify for the fourth point because researchers have found that a majority of inmates meet criteria for antisocial personality disorder (Cunningham & Reidy, 1999; Hare, 2003). Hence, many criminal forensic patients meet two criteria in the *DSM-IV* purely by default. Regarding misclassification, Rogers (1990) conducted a study exploring the accuracy of *DSM-IV* indices in identifying malingerers. This study found that found a very high false positive rate (79.9 to 86.4%), indicating that the vast majority of suspected malingerers were, in actuality, miscategorized. In fact, the *DSM-IV* indices accurately identified malingerers only 13.6 to 20.1% of the time.

Such research findings should prompt professionals to apply DSM-IV indices very cautiously, especially when dealing with forensic populations. It is, perhaps, most advisable to treat them as screening criteria, using them to prompt a more thorough evaluation. Historically, there has been much debate as to the most appropriate way for mental health professionals to conduct a thorough assessment of malingering. What follows is an account of how such assessment methods have evolved within the field of psychology. The subsequent section delineates how methods for the detection of malingering began with simple observations of case studies in the early nineteenth century. Then, in the century that followed, procedures to uncover feigning gradually evolved from case observations, which yielded little to no reliable guidelines, to the use of standardized assessment measures employing detection strategies that are

conceptually-based and empirically-validated for use with various groups of people (Rogers, 1997; Wessely, 2003).

Early Methods of Detecting Malingering

In their review of detection procedures used in the nineteenth century, Geller, Erlen, Kaye, and Fisher (1990) describe various warning signs commonly believed to indicate feigned disorders throughout the 19th century. It seems that many of these early detection methods revolved around a mental health professional's ability to recognize "signs" exhibited by the typical malingerer. Such signs include: (a) specific interview behavior (e.g., inability to maintain eye contact and hesitation in responding), (b) feigned presentation (e.g., symptoms increase while being observed, and overacting), (c) areas of intact functioning not usually observed in genuine patients (e.g., no sleep disturbances, and no appetite disturbance), and (d) atypical symptoms (e.g., rapid onset, overly absurd thoughts, no fluctuation of symptoms, and decompensation does not follow typical patterns). Indeed, references are even made to ancient papyrus writings from 900 B. C. that describe behaviors, mannerisms, and other presentations common among deceitful individuals (Resnick, 1984). These methods imply that for evaluators to accurately detect dissimulation, they must become adept at spotting every possible "sign."

Early Interventions for Malingering

Just as the early detection methods lacked systematic protocols, so did early methods of confirming a classification of malingering. Interventions relied largely on unstandardized clinical tactics. For example, some professionals in the field attempted to use questioning and observation to expose areas of intact functioning that were divergent from purported impairment.

Others even advocated using coercive methods to prompt the individual to confess (Geller et al., 1990). Success using these methods was far from guaranteed.

Due to the limitations of case studies, initial methods used to establish malingering in the early twentieth century had unknown validity and reliability. Nevertheless, the clinical judgment and professional expertise of physicians as well as mental health practitioners were heavily criticized in cases where malingering went undetected. Doctors were judged as careless for not closely evaluating and carefully documenting every possible "sign" of dissimulation, and it was widely assumed that a more qualified physician would not have committed such an error (Pope, 1919). Interestingly, however, there was also an awareness that malingering was difficult to detect and determinations were quite subject to evaluator bias. While Pope (1919) suggested that evaluators remain suspicious of their clients and scrutinize responses for any sign of an ulterior motive, Meagher (1919) emphasized, that "a suspicious mind can discover almost anything to corroborate its suspicions" (p. 966) and evaluators' increased scrutiny did not necessarily yield more accurate results.

In the end, case studies are useful for documenting specific instances of malingering and exploring its characteristics within the context of that particular case. However, specific characteristics cannot simply be generalized to other cases, leaving the clinical usefulness of case studies to be very limited. It is impossible to establish standard criteria for the determination of feigning by becoming familiar with the particulars of any single salient case or even a vast number of individual cases, as was suggested by Pope (1919).

Some modern day guidelines for clinical decision making parallel the spirit of these early clinical methods, as well as their criticisms. Specifically, hypothesis-testing *models* encourage evaluators to form opinions early on in an evaluation and gather assessment data that will either

prove or disprove these initial hypotheses. Just as Meagher did in 1919, modern researchers (Borum, Otto, & Golding 1993; DeClue, 2002) point out that these approaches might create bias if evaluators overly commit to their initial hypotheses and fail to fully test alternatives. The inception of well-researched standardized assessment measures reduced a great deal of the bias inherent in unstructured early interviews. A discussion of how they changed the face of malingering assessment will soon follow. First, however, it is important to discuss the theoretical framework upon which these validity scales are based: *detection strategies* for malingering.

The Development of Detection Strategies

Detection strategies are standardized, theoretically based methods that have been empirically tested and validated for differentiating between specific response styles used in standardized assessment measures (Rogers, 1997). To be established as valid, a detection strategy must be researched and tested by multiple scales on different test measures. Multi-method systems of validation emphasize the importance of large effect sizes for the accurate classification of feigners and genuinely impaired individuals. The introduction of standardized assessment measures made this thorough testing of detection strategies possible.

In 1997, Rogers described a number of detection strategies for feigned psychopathology. These strategies classify different domains of feigning. They have been validated through both original research and replication studies, and have been tested through the use of more than one method of assessment (e.g., interviews, and multi-scale inventories). Rogers' ten accepted detection strategies for feigned psychopathology are summarized below in Table 1.

Table 1

Description of Detection Strategies

Detection Strategy	Overview
Rare symptoms	Focuses on symptoms that rarely occur in psychiatric patients; over-endorsement of uncommon symptoms implies that the client is exaggerating or feigning.
Improbable symptoms	Focuses on the number of symptoms endorsed by a person that are so outlandish, they are highly unlikely to be true symptoms of a disorder. There is increased reason to question the person's account when high numbers of improbable symptoms are reported.
Symptom combinations	Focuses on inquiries about true psychological symptoms. However, some unusual symptom pairs are rarely observed in genuine patients. Over-endorsement of rare combinations implies malingering.
Obvious symptoms	Focuses on whether the person being evaluated reports a larger-than-expected number of symptoms that are clear indicators of psychopathology.
Subtle symptoms	Focuses on whether the person being evaluated endorses relatively few symptoms seen as common difficulties not necessarily indicative of mental disorders.
Symptom selectivity	Focuses on how selective examinees are in their endorsement of psychological problems. Malingerers tend to endorse a wider array of symptoms from various disorders than genuine patients typically do.
Symptom severity	Focuses on how the person being evaluated characterizes the intensity of their symptoms. Genuine patients will typically identify some of their symptoms as being worse than others. However, malingerers tend claim that many of their symptoms are "extreme."
Reported vs. observed symptoms	Focuses on the clinician's own observations compared to the symptoms that the client reports. When the client reports a much higher number, it may be because the person is reporting false symptoms.
Spurious patterns	Focuses on patterns of response that are characteristic of malingering, but are very uncommon in clinical populations.
Erroneous stereotypes	Focuses on whether the person being evaluated reports an excessive number of misconceptions about mental disorders held by the general population. If so, the issue of feigning is raised, as people who do not actually suffer from a particular disorder may be misinformed about symptoms and their presentation.

In understanding the application of these strategies, Miller's work (2001) provides a useful illustration in creating a malingering screen, the Miller Forensic Assessment of Symptoms Test (M-FAST; Miller, 2001). The M-FAST included scales to assess the following detection

strategies in her measure: Reported vs. Observed (RO), Extreme Symptomatology (ES), Rare Combinations (RC), Unusual Hallucinations (UH), Unusual Symptom Course (USC), Negative Image (NI), and Suggestibility (S). Four of the scales are similar to detection strategies identified by Rogers (1997); they include rare symptoms (UH), symptom combinations (RC), reported vs. observed (RO), and severity of symptoms (ES). These strategies rely on unlikely presentation of symptoms (Vitacco, Jackson, Rogers, Neumann, Miller, & Gabel, 2008). Combining these detection strategies, research generally finds that the M-FAST is a valid screen for the detection of feigned psychopathology (Guy, Kwartner, & Miller, 2006; Jackson, Rogers, & Sewell, 2005).

These and other detection strategies have been extensively tested in two separate formats: multiscale inventories and structured interviews. As noted by, Guy et al. (2006), multiscale inventories focus on embedded validity scales. The second approach moves away from embedded indicators and examines specialized measures of malingering. Each approach has its shortcomings (Rogers & Bender, 2003). To illustrate their limitations, both types of assessment measures are discussed and a comparison of effect sizes across detection strategies and assessment measures follows the appraisal of multiscale inventories below (see Table 2).

The Impact of Standardized Assessment Measures on Feigning Research

Feigning is notoriously difficult to detect by clinical interview alone and even experienced mental health professionals are often unsuccessful. Early research (Bourg, Connor, & Landis, 1995) reveals that clinicians conducting interviews of examinees are generally poor evaluators of malingering. This lack of success is likely due to the fact that clinical interviews are not standardized and rely almost exclusively on the mental health professional's own judgment. Problems with the validity of unstandardized clinical evaluations have been noted

throughout the history of assessment (Borum, Otto, & Golding 1993; DeClue, 2002; Geller et al., 1990; Meagher, 1919; Pope, 1919; Resnick, 1984). When clinicians do not perceive the client's deceptive intent, or when they do not make sufficient inquiries, malingering can go undetected (Rogers, 1997; Rogers & Schuman, 2005). However, clinicians with no forensic training and little previous exposure to true feigners can be, in fact, very accurate evaluators of malingering when they are provided with information in addition to client interview data, such as a client's personal history and scores from standardized assessment measures (Bourg et al., 1995). Thus, valid testing measures such as multiscale inventories and structured interviews are crucial.

Assessment of Feigning Using Multiscale Inventories

It was the advent and growing sophistication of multiscale inventories that first allowed for a method of systematically evaluating feigners and non-feigners through comparing differences between the two groups. The original MMPI (Hathaway & McKinley, 1940) vastly changed the assessment of response styles and malingering when Meehl and Hathaway (1946) noted that a clinician must assume that patients might be motivated to alter their symptom presentation on purpose. These early researchers found it of utmost importance to include scales to assess response style in order to determine the genuineness of a client's self-report. The MMPI and other multi-scale inventories are discussed in further detail within the following sections.

MMPI-2

The Minnesota Multiphasic Personality Inventory – 2 (MMPI-2; Butcher et al., 1989) is among the most widely researched multiscale inventories that includes well-established validity scales. The original MMPI was one of the first personality tests to incorporate validity scales into

the interpretation of its results. Basic MMPI-2 validity scales include those designed to determine whether one is responding in a defensive manner, inconsistent manner, or over-reporting symptoms of severe psychopathology (Greene, 2000).

In a meta-analysis of the MMPI-2 and malingering, Rogers, Sewell, Martin and Vitacco (2003) reviewed detection strategies. One main focus of the MMPI-2 is "quasi-rare" strategies such as those found on the F and Fb scales. The term "quasi-rare" signifies that the items are uncommon within normative samples, but not necessarily among genuine clinical patients. Rogers and Bender (2003) cautioned against relying exclusively on F-scale elevations because true patients with severe psychotic disorders might be misclassified. Specifically, a high score on the F-scale is not necessarily indicative of malingering; instead, it can mean that the person is responding honestly and exhibits genuine, albeit uncommon, symptoms such as those found in schizophrenia. In fact, Rogers et al. (2003) found patients with schizophrenia had marked elevations on F (M = 80.1). Surprisingly, this finding is far from new; Gough (1947) found that genuinely psychotic patients typically score in the elevated range on the MMPI F scale (M = 75.38).

The F scale was not initially intended to be a measure of malingering. It is composed of 64 items uncommon in a non-clinical normative sample. As a result, it is simply an indicator of atypical responding that could be a result of anything from confusion, reading difficulties, and a person's pathological construal of their own life experiences (Meehl & Hathaway, 1946, p. 536).

PAI

The Personality Assessment Inventory (PAI) (Morey, 1991; Morey, 2007) is a newer-generation multiscale personality measure that uses validity scales to identify response styles

including malingering. Of the PAI validity scales, Negative Impression Management (NIM) is most often used to assess malingering, as high NIM scores likely suggest that the participant is exaggerating symptoms or endorsing a large amount of extremely bizarre symptoms. The NIM scale employs a *rare symptoms* strategy (Rogers & Schuman, 2005). Beyond NIM, the Malingering Index (MAL) uses the spurious patterns strategy to examine configural rules indicative of feigned mental disorders.

One major advantage in using the PAI is that, unlike the MMPI-2, its validity scales do not overlap with symptoms included on the clinical scales. Thus, the PAI does not suffer from the same problems as the MMPI-2's F scale, where genuine rare symptoms might confound the assessment of malingering and cause honest responders to be misclassified. Despite this advantage, Rogers and Bender (2003) caution that the PAI should not be used as the sole measure used to detect malingering because only extreme elevations on NIM and MAL are indicative of malingering. Less extreme elevations can also be found among patients with genuine disorders. For this reason, the PAI could be used to rule out persons not likely to be feigning, but a more rigorous evaluation, using multiple measures should be used before an evaluator characterizes a person as a malingerer. Though results of their meta analysis indicate that feigning indices on the PAI were strong predictors of both coached and uncoached malingering, Hawes and Boccaccini (2008) caution that the limited number of feigning studies prevents detailed analysis and conclusions.

Summary

The research studies on the MMPI-2 and PAI cited above call for mental health professionals to proceed with caution when employing these measures in an assessment of

feigning. Indeed, experts in the field differ as to whether they believe structured interviews or self-report measures are better for the assessment process. On the one hand, structured interviews are time and labor intensive. They require the services of a clinician with considerable training in their administration, are often times lengthy, and can only be administered to one patient at a time. This process ties up quite a bit of resources and may not be the most desirable option for mental health clinics. Self-report measures, on the other hand, can be administered quite easily to an entire group of patents at one time (Guy, Poythress, Douglas, Skeem, & Edens, 2008). Aside from the aforementioned classification problems caused when clinical symptom scales overlap with validity scales, generalized self-report measures also minimize the opportunity for clinicians to observe behavior during the assessment, and clarify questions to obtain the most useful information from patients (Edens, Poythress, & Watkins-Clay, 2007).

Moving to the second line of research described by Guy et al. (2006), the following section addresses the utility of specialized malingering measures. Following this, effect sizes to those of specialized measures are compared.

Specialized Measures of Malingering

The M-FAST, a structured interview intended to screen for malingering (Miller, 2001), was previously discussed (see p. 11). An important benefit to using a structured interview rather than an unstructured clinical interview to diagnose a mental disorder or determine feigning is that it ensures the interviewer systematically assesses specific domains with a standardized recording of response styles and other clinical complaints. In an unstructured interview, the evaluator might inadvertently omit questions that could be important for the final determination simply because they did not think to ask (Rogers, 2001). The following section addresses a

comprehensive malingering measure that investigates the construct more thoroughly than most screening measures.

SIRS

The Structured Interview of Reported Symptoms (SIRS; Rogers et al., 1992) is a comprehensive measure that was designed to evaluate feigned mental disorders. This structured interview is widely considered the gold standard for the detection of feigned mental disorders (Rogers, 2001; Rogers, 2008). In a survey of forensic diplomates conducted by Lally (2003), the majority of respondents recommended the SIRS for use in the evaluation of malingering. Especially in forensic contexts, the SIRS is the most researched specialized measure for the assessment of feigning. One criticism of the SIRS is that it often requires a lengthy administration time (Green, Rosenfeld, Dole, Pivovarova, & Zapf, 2008). Despite this critique, the SIRS is commonly chosen in evaluations because of its high reliability and validity (Blau, 1998).

The SIRS is composed of 172 items that are organized into 8 primary scales: Rare Symptoms, Symptom Combinations, Improbable and Absurd Symptoms, Blatant Symptoms, Subtle Symptoms, Selectivity of Symptoms, Severity of Symptoms, and Reported vs. Observed Symptoms. Each primary scale employs a different detection strategy; these detection strategies and additional strategies described by Rogers (1984) are summarized in Table 1. For most of the questions, clients' responses are quantified as "no," "qualified yes," (e.g., "sometimes") or "yes." Additionally, 32 items are asked twice within the same SIRS administration and the two responses are compared to gauge response consistency (Rogers et al., 1992).

Blau's (1998) review finds that the SIRS has a high success rate in identifying malingerers in competency-to-stand-trial evaluations; it is also successful in other clinical and forensic contexts. The SIRS is organized into two general domains: "Unlikely Presentation" (i.e., implausible symptoms unlikely to be true) and "Plausible Presentation" (i.e., potentially genuine symptoms but reported to an inordinate degree). Research findings indicate that the SIRS is an effective measure using the two domains of detection strategies based on spurious and plausible presentations (Rogers, Jackson, Sewell, & Salekin, 2005).

The following table compares the usefulness of detection strategies espoused by different standardized assessment measures by evaluating their effect sizes. Using Rogers' (2008) guidelines, effect sizes greater than 1.25 are considered large and effect sizes greater than 1.50 are considered very large.

Table 2

Effect Sizes (Cohen's d) for Detection Strategies used in Measures with Feigning Indices

Detection Strategy	SIRS	MMPI-2	PAI	M-FAST
Rare symptoms	RS = 2.15	Fp = 1.90	NIM = 1.39	UH = 1.06
Improbable symptoms	IA = 1.87			
Symptom combinations	SC = 1.84			RC = 1.35
Obvious symptoms	BL = 2.21	Obv = 2.03		
Subtle symptoms	SU = 1.54	Su = .68		
Symptom selectivity	SEL = 1.79			
Symptom severity	SEV = 1.89			ES = 1.56
Reported vs. observed symptoms	RO = 1.82			RO = .80
Spurious patterns			MAL = 1.11 RDF = 1.84	
Erroneous stereotypes		DS = 1.62 FBS = .32		

Note. This chart uses data compiled from Rogers et al. (2003), Sellbom and Bagby (2008), Rogers (2008), and Jackson et al. (2005).

In comparing the number of detection strategies employed by standardized assessment measures, it is evident that all use the rare symptoms strategy. Effect size for the SIRS primary rare symptoms scale is larger (d = 2.15) than that of the MMPI-2 (d = 1.90), the PAI (d = 1.39), and the M-FAST (d = 1.06). Other than rare symptoms strategies, there is no complete overlap among detection strategies for the MMPI-2, PAI, and M-FAST, thus there is no systematic way to compare their efficacy. However, effect sizes for the SIRS scales are larger than each of the corresponding scales on the other three measures. Additionally, the SIRS employs the most detection strategies out of these measures and effect sizes for each are considered large to very large. These results help to solidify its status as the gold standard in terms of measures for the determination of malingering; its status provides the primary rationale for the current study. A chief goal of this research is to validate a Spanish language version of the gold standard in malingering detection and make it available to a growing monolingual Hispanic population that currently has no access to other specialized measures of malingering.

The Growing Need for Spanish-Language Assessment Measures

Psychologists and other mental health professionals are aware that most standardized assessment measures were developed for clients proficient in English and subsequently normed on samples comprised mainly of European American individuals. However, contemporary methods of psychological assessment in the United States have begun to face unique challenges in a rapidly changing cultural landscape with increased diversity among the populations needing mental health interventions. For instance, the Hispanic population is currently the fastest-growing minority group in the United States. According to the most recent available census data, approximately 40% of the US Hispanic population is foreign-born and may primarily speak

Spanish (US Census Bureau, 2004). In fact, of the general US population that reports speaking English less than "very well," nearly 56% predominantly uses the Spanish language (US Census Bureau, 2000). This growing subpopulation creates a need for assessment tools that are reliable and valid for use with Spanish-speaking Hispanic populations.

Language plays an increasingly important role in test validity because there is a growing segment of the United States for whom traditional measures in the English language cannot be effectively used (Solano-Flores, Backhoff, & Contreras-Niño, 2009). To date, a small number of Spanish-language measures have been properly validated. These measures mainly include multiscale inventories whose English language versions are widely used in research and clinical practice. Particular examples include the Spanish MMPI-2 (Lucio, Reyes-Lagunes, & Scott, 1994) and the PAI (Morey, 2007). However, there remains a wide variety of important psychological measures that have yet to be approved for use with this growing minority population. A new Spanish language version of the SIRS has been developed, but its psychometric properties have not been rigorously evaluated. Validation of such a measure is the chief goal of the current study because, to date, no measures for the detection of feigning in clinical and forensic settings have been approved for use with Spanish-speaking populations.

Test Bias and Validation of Test Translations

The main reason it is imperative to validate a translated measure is that psychometric characteristics for standardized assessment measures change when they are administered to individuals who are culturally different from the normative sample (Geisinger, 1994; Marin & Marin, 1991), especially when the language of the test items also changes. Ethical guidelines from the American Psychological Association require that psychologists working with

ethnically, linguistically, and culturally diverse populations should recognize these characteristics as important factors affecting a person's experiences, attitudes, and psychological presentation (Bersoff, 2004). Differences between the test scores of individuals from a minority group and those from the dominant culture can become problematic when they lead to inaccurate predictions or misdiagnoses for minority individuals (Graham, 1990). Thus, only tests that have been formally translated into Spanish and subsequently validated should be made available for use in clinical practice.

For assessment measures, many areas of potential bias are discussed in the test translation literature. Many areas affecting the validity of assessment measures used with ethnic minority populations revolve around the *etic* and *emic* qualities of the test (Dana, 1993, 2005; Olmedo, 1981). Etic measures are those with "universal" applications, whose constructs are equally applicable to individuals of all different groups. It is expected that an individual's assessment results on an etic measure can be interpreted based on the same set of norms, regardless of the individual's membership in any particular cultural group. Emic measures, on the other hand, are culture-specific measures; their clinical applications can be specific to populations based on age, gender, ethnicity, or any other grouping classification. It is understood that emic measures are only appropriate for use with the groups for whom they were designed.

Researchers (Berry 1969, 1988, 1989; Dana 2005) have observed for some time that most standardized assessment measures are normed on samples comprised mainly of European Americans. Based on current clinical practices, they fall into the category of *imposed etic* tests. This is to say, interpretive norms were developed mostly on individuals of European American heritage, without further testing on other cultures, and remain valid for only the European American culture. Without validation studies to establish culturally relevant cut scores and

interpretation guidelines, test developers imply that European American based cut scores are universally valid and generalize to all cultures. This serious omission in test development effectively forces culturally different individuals into the same interpretative categories as European Americans, thereby creating a substantial possibility for misdiagnosis and misinterpretation of test results (Dana, 1993; Todd, 2005).

Though they do not use the same labels, Van de Vijver, and Hambleton, (1996) point out areas that seem to highlight the emic qualities of a test when discussing their three domains of potential test bias. They suggest scrutinizing each test for the presence of: *construct bias*, or whether the construct actually applies to both cultural groups, *method bias*, and *item bias*. Of the latter terms, method bias often relates to method of administration, such as asking a cultural group to perform a task with which they are unfamiliar. Common examples include: Likert-type scales, multiple choice tasks, and other activities which are not routinely encountered in the school systems of some countries. Lastly, *item bias* refers to questions that contain poor wording, offensive material, or other abnormalities that affect how a person responds. Minimizing these biases in an assessment measure will, likely, minimize any *imposed etic* effects on the cultural group being assessed.

Guidelines for Test Validation

Hambleton (2001) summarized the Test Translation and Adaptation Guidelines developed by the International Test Commission (ITC) in 1992. As stated in the ITC Test Translation and Adaptation Guidelines, when a test is adapted into a new language, test developers and publishers must apply appropriate research methods and statistical techniques to establish the validity of a test in each population for whom the adapted version is intended.

This guideline requires test developers to use research results to improve the accuracy of the translation/adaptation process and identify problems in the adaptation that may render a measure inadequate for use with the intended populations. The uncovered problems should be noted and remedied, if at all possible. Second, test developers should strive to establish the equivalence of the different language versions of the test, to make them as parallel to the original as possible. Third, the validity of the translated version must be determined separately from that of the original measure. It should not be assumed that a translated version has acceptable validity simply because that of the original English language version is adequate (Allalouf, 2003; Anastasi, 1988).

Until the reliability and validity of these assessment measures has been determined, mental health professionals should refrain from using them just as they would refrain from administering any other unvalidated measure (Allalouf, 2003; Hambleton, 2001). For this reason, researchers have long criticized translations of multi-scale personality inventories for being made available to clinicians before sufficient validation studies have been conducted. The danger in administering tests that have not been validated is that clinicians interpret the results based on an *assumption* that test continues to function in the intended manner (Fantoni-Salvador, 1997; Rogers, Flores, Ustad, & Sewell, 1995).

Challenges to the validity of translations address different facets of equivalence such as, *linguistic equivalence, functional equivalence* (applicability of the same construct to the new cultural group)*, cultural equivalence,* and *metric equivalence* (e.g., similar item difficulty; Peña, 2007). In a study evaluating the McCarthy Scales of Children's Abilities (MSCA), Valencia and Rankin (1985) found evidence of *content bias* in the Spanish translated measure. When controlling for total score, they discovered that children who took the test in Spanish performed

substantially lower on two specific scales. These differences were attributed to language and cultural differences in the Spanish-speaking group of children. Thus, their lower scores on those two scales are attributable to the emic qualities of the test.

Development of Valid Spanish-Language Measures

Clearly, the process of creating useful Spanish language measures is complex and requires more than simple translation of the original measures. Even when measures are meticulously translated, perfect linguistic equivalence is often impossible to achieve because some languages have fundamental differences in the way they are structured. Dissimilarities create differences of varying magnitude in the meaning of test items and in the types of responses that these items elicit (Anastasi, 1988; Solano-Flores, Backhoff, & Contreras-Niño, 2009).

Translated phrases can often sound stilted, awkward, and confusing in the new language (Peña, 2007). Also, meaning of the new test items is highly dependent on the translator's word choice. When a particular word is chosen in place of a related synonym, the meaning of the item could be changed in subtle ways. A far more serious problem occurs when researchers strive for a completely literal translation under the mistaken belief that the words will continue to express the same construct in the new language that was expressed in English. At times, a direct translation must be sacrificed so that test items can more accurately convey the intended *message* (Jeanrie & Bertrand, 1999; Marin & Marin, 1991; Solano-Flores, Backhoff, & Contreras-Niño, 2009; Van de Vijver, & Hambleton, 1996). For example, idiomatic English language phrases whose meanings are well known in the United States, such as "quitting cold turkey," or "he has the blues" have no corresponding meaning when translated directly into Spanish. In this

situation, a word-for-word translation would be completely ineffective. Instead, translators must concentrate on the ideas that these phrases represent and communicate them in words that will be understood in a different language and by a different culture.

Variations in how groups with different cultural or ethnic backgrounds tend to respond also impact the efficacy of a measure. Thus, differences in response patterns of distinct ethnic groups must be empirically researched so that they can be taken into account when interpreting the measure (Anastasi, 1988).

In addition to the different response patterns among distinct cultures, level of acculturation for the members of each ethnic group should be assessed. Acculturation can be defined as the changes that occur in an individual's beliefs and behaviors, as a result of interaction with his own ethnic group (e.g., Hispanic) and another cultural group (e.g., European American). Individuals with higher levels of acculturation have a greater understanding of the new culture (American) and begin to accept and incorporate aspects of it into their daily lives. Individuals with low levels of acculturation will continue to chiefly identify with the values of their ethnic group despite interaction with members of a different culture (Wagner & Gartner, 1997). In 1989, Berry et al. proposed a two-dimensional model of acculturation. In this model, individuals feel a need to identify with both their own minority culture and with the majority culture. The individual can maintain one of four possible relationships with majority and minority cultures:

- Assimilation: sole identification with the majority culture
- Integration: identification with both cultures
- Separation: sole identification with the minority culture
- Marginality: no identification with either culture

Berry's (1989) is a bidimensional model of acculturation, where it is possible for the individual to maintain varying degrees of affiliation with minority and mainstream cultures. In contrast, there are also unidimensional models of acculturation, which contend that one relationship must always be stronger than the other (Gordon, 1964). In unidimensional models, individuals relinquish their ethnic culture, as they become more assimilated to mainstream American culture.

In both models, distinct levels of acculturation augment the variety of possible response patterns because differences also exist within cultures, not just between them, depending on how much an individual identifies with each of the cultures in question. Unidimensional models might obscure the complexity of individual acculturation, by failing to recognize bicultural individuals who identify strongly with both cultures (Ryder, Alden, & Paulhus, 2000). However, both models emphasize the notion that all members of an ethnic minority cannot be grouped together when data are analyzed. How acculturation affects responses to test items should also be established when characterizing new normative samples and cut scores.

Culturally-Specific Response Patterns Common Among Hispanic Americans

Culturally-specific response patterns identified in the literature affect the validity of psychological assessments—whether they are conducted in English or Spanish—and should be taken into account when interpreting assessment results in order to avoid misdiagnosis due to imposed etics. Thus, what follows is a discussion of response patterns commonly displayed by Hispanic American individuals on various standardized assessment measures.

In a classic study, Molina and Franco (1986) found significant differences in self-disclosure based on both ethnicity and gender. Overall, Mexican Americans tended to self-disclose less than their European American counterparts. Moreover, Mexican American men

self-disclosed even less than Mexican American women. It is imperative that clinicians are aware of cultural response patterns. If individuals from a different cultural background, such as Latino, appear to respond in a guarded or defensive manner during assessment, this observation can have a significant impact on the validity of their clinical profiles and the subsequent accuracy of their diagnoses (Helms, 1992). Research conducted with Hispanic individuals and the MMPI suggests results similar to Molina and Franco are found on multiscale inventories. In an early review by Campos (1989), several studies consistently found significant "L" scale elevations among Hispanic Americans when compared to European Americans. Similar results have been found for Hispanic American women on the MMPI-2 (Callahan, 1998). The most logical conclusion is that Hispanic Americans, consistent with their culture, are reticent to disclose their psychological issues in the formal context of an evaluation. This reticence to express feelings can best be described as a desire for privacy and selectivity about with whom personal problems can be shared. This ethnically sensitive interpretation is different from a defensive response style—the common interpretation used for European Americans.

Rogers and his colleagues (Fantoni-Salvador & Rogers, 1997; Rogers, Flores, Ustad, & Sewell, 1995) conducted several studies on the clinical usefulness of the Spanish-language PAI.One omission from the Fantoni-Salvador and Rogers (1997) study was an examination of PAI validity indicators that could address the previously described issue regarding a reticence of Hispanic Americans to disclose as much as European Americans. A more recent study by Hopwood, Flato, Ambwani, Garland, and Morey (2009) looked closely at the PAI and socially desirable response styles in Hispanic Americans. Using undergraduate students, Hispanic American participants attained higher scores than European Americans on all socially desirable response measures used in the study, with statistically significant differences for the following:

PAI Defensiveness indicator (DEF), PAI Cashel Discriminant Function (CDF), and the Marlowe-Crowne Social Desirability Scale (MC). However, these differences produced only modest effect sizes, (*ds* = .28, .37, and .38, respectively).

A dissertation by Romain (2000) found that more than 40% of the PAI protocols from Hispanic Americans were considered "invalid" based on the standard cut scores outlined in the PAI manual (Morey, 1991), as compared to 20% of the European American profiles. Unfortunately, she did not provide specific data about response styles so it is unclear what are the proportions that were potentially defensive, feigned, or highly inconsistent. Some differences for Hispanic Americans may be attributable to acculturation: 45% of the monolingual versus 37% of the bilingual individuals had invalid profiles. Although Hispanic Americans had substantially higher "Positive Impression Management (PIM)" scores in comparison to European Americans (Cohen's *d* = .60), both groups evidence relatively little defensiveness with mean PIM scores of 45.32 and 38.06 respectively (see Romain, 2000). Her data also suggest Hispanic Americans are scoring in a non-normative or atypical manner on items unrelated to psychopathology (i.e., INF scale). While INF elevations may reflect carelessness, confusion or reading difficulties, psychologists may wish to consider issues of reading comprehension and acculturation before considering these interpretations. Given its large effect size (*d* = 1.00), INF scale may indicate a culturally specific response pattern beyond differences in reading abilities. Elevations in scales that evaluate defensiveness and socially desirable responding raise the possibility that Hispanic Americans are potentially reticent to disclose treatment issues related to psychopathology and support from others. Not to overpathologize minority clients, the alternative explanation is that these issues are less salient to Hispanic American clients.

Translation Techniques

The test translation process has been equated to construction of a new test, requiring evidence for construct validity, statistical support, and assessment of bias at the item level (Jeanrie & Bertrand, 1999). Test developers must be prepared to provide each of these requisite pieces of information for a valid measure, and should consider their ability to do so when choosing among the translation approaches described below.

Three basic approaches are generally used in translating written documents from one language to another: one-way translation, translation by committee, and back translation (Marin & Marin, 1991). Each technique varies in complexity and has its own set of strengths and limitations. The following section aims to describe these commonly used methods.

One-Way Translations

One-way translations employ the simplest of translation techniques. Here, one bilingual individual uses dictionaries, reference materials, and his or her knowledge of both languages to create the translated product (Marin & Marin, 1991). This approach is appealing because it is time-efficient and cost-effective, using the resources of only one person to achieve a translation. However, its simplicity is also the basis of most criticisms. Relying on a single person to translate the material leaves the translation vulnerable to error, because the translator's work is left unchecked. When any misinterpretations make their way into the final product, quality of the translated measure is adversely affected. Berkanovic (1980) demonstrated that a health survey with one-way translation into Spanish had different psychometric properties and lower reliability than its English language counterpart.

One recommendation is to focus on the quality of the translator to improve the quality of

translation. Hambleton and Kanjee (1995) stress that translators should be (a) highly proficient in both languages involved, and (b) familiar with cultural groups associated with both languages. The latter recommendation would help in constructing translated items that flow well in the new language, retain the intended meanings, and are understood by the target population. If translators also have an understanding of the construct being measured, they will better be able to preserve the intended meaning of test items. Despite these suggestions to improve the process of one-way translation, researchers (Brislin, 1970; Marin & Marin, 1991) tend to agree that it should not be used. Instead, they conclude that more translators should be involved, and that back translation should be used for quality control.

One-way translations can be made more thorough via the use of judges to evaluate the final product (Jeanrie & Bertrand, 1999). Judges can evaluate the following three areas: *content equivalence* (i.e., relevance to that cultural group), *conceptual equivalence* (i.e., maintaining construct validity), and *linguistic equivalence* (i.e., maintaining as direct a translation as possible without jeopardizing content and conceptual equivalence). The items attaining the highest scores can be compiled and edited in the final step. This framework attempts to remedy some of the major critiques of one-way translation, but with no data provided by the researchers, it is impossible to determine if it leads to a better one-way translation or simply takes up resources that could best be put to use in implementing a more well-researched translation model.

Translation by Committee

A second translation technique is that of translation by committee (Marin & Marin, 1991). This approach utilizes two or more individuals who are familiar with both languages. Each individual produces their own translation without consulting the other translators. After the

initial translations are complete, the person who commissioned the translations can ask all translators to meet, compare their individual versions, discuss, and resolve the differences. In this manner, they create a final version incorporating the changes they have discussed. The goal of this process is to prevent problems, such as misinterpretation and awkward wording that arise from relying too heavily on a single translator. Alternatively, the person who commissioned the translations can ask one more persons (not involved in the original translations) to review each translator's work and choose the best version (Marin & Marin, 1991). This option still falls under the rubric of translation by committee because there are multiple translators involved in the process.

Translation by committee can be more accurate than one-way translation. Marin and Marin (1991) are quick to point out, however, that traits shared by the translators such as cultural background, education level, and social class, might lead them to make the same errors in their independent translations. Ensuring that the committee consists of individuals with diverse cultural backgrounds (Spanish, American, Mexican, Puerto Rican, Peruvian, etc.) reduces the risk of error caused by uniform interpretations (Martinez, Marin, & Schoua-Glusberg, 2006). However, the committee discussion can never ensure that all possible mistakes are pointed out, as committee members might not feel comfortable criticizing each other's translation (Marin & Marin, 1991).

Back-Translation

A final translation procedure, commonly known as "back translation," is the most recommended by researchers (Brislin, 1986; Moreland, 1996), yet it remains the least used translation technique (Jeanrie & Bertrand, 1999). It's lack of use may be because of its time-

consuming nature. Back-translation makes use of at least two bilingual individuals. As in one-way translation, one of the individuals independently translates the original language (e.g., English) into the new language (e.g., Spanish). At this point, a second translator takes the newly translated version and translates it back into the original language. Of critical importance, the translators must work independently throughout this process and are not permitted to consult one another. There are now two English language versions of the measure: the original version and the back-translated version. The two English versions are compared and inconsistencies are identified. When differences are found it is critical to approach both translators, determine why the difference exists, and reach an agreement about the best option (Marin & Marin, 1991). A third party not involved in the original translation or back translation can also be commissioned to evaluate the two English versions (Brislin, 1970).

Back translation can be improved if the process is conducted more than once and with additional translators being used. These iterations make the procedure more time-consuming and complex. However, the measure is reviewed by more bilingual professionals, which produces a better version of the instrument in the end (Marin & Marin, 1991). Back translation has been used extensively in creating Spanish language versions of assessment tools as diverse as general health questionnaires (Marin, Perez-Stable, & Marin, 1989), and structured interviews for the diagnosis of mental health problems (Burnam et al., 1983).

Marin et al. (1989) advocated the back translation process, finding that the Spanish version of their survey was, indeed, equivalent to the English version after administering both versions to bilingual speakers. Sireci, Yang, Harter, and Ehrlich (2006) conducted a study designed to determine how a more rigorous translation procedure (back translation) compared to a simple translation procedure (one-way translation). They found that for many of the test items,

back translation produced results that were more comparable to the original English measure. Using their design, the Spanish DIS was also found to be acceptably equivalent to the English DIS for bilingual participants (Burnam et al., 1983).

Back translation is the preferred method for most researchers (Marin & Marin, 1991). However, it is only useful when those translators performing the back translation stay as close as possible to the version with which they are working. It is easy for a good translator to automatically fix problems they encounter so that their final product is easy to understand. When they fix errors in the Spanish version while translating it back to English, problems left in the Spanish version are never discovered. Translators must, thus, be advised to resist the urge to correct mistakes they encounter, no matter how small they may seem. It is best to translate these errors and bring them up as points of discussion when the new versions are reviewed and edited.

A limitation in the process of back translation is that it still relies on the translator's interpretation of item meaning (Marin & Marin, 1991). For this reason, it is important to employ the same precautions that should be used for "translation by committee." Recruiting translators from varied educational, cultural, and social backgrounds minimizes errors caused by uniform interpretations (Martinez, Marin, & Schoua-Glusberg, 2006). Another criticism of back-translation is that it provides no guidelines as to how many independent translators are sufficient for a good translation (Cha, Kim, & Erlen, 2007). Some experts, instead, advocate using a combined technique (Jones et al., 2001) that employs back-translation *and* administers both versions of the test to bilingual participants in order to identify discrepancies before creating the final version. This method appears to incorporate equivalence testing (a recommended step for final validation) into the translation procedure (Hambleton, 2001).

The Spanish SIRS

The Spanish SIRS was translated from the original English-language version of the SIRS through the method of back translation. Specifically, the Spanish version was created using a multi-step and thorough translation method where three bilingual psychologists made an initial translation from English to Spanish. Each translator made an independent translation and then met to discuss and compare the differences. From this comparison, a working translation was developed. It was then back-translated independently by yet another bilingual psychologist who had no knowledge of the original English-language version. A fifth bilingual psychologist evaluated the original English version and the back-translated version, found discrepancies, and made corrections, producing the final translation.

Purpose of the Current Study

This study sought to determine whether the Spanish SIRS effectively distinguishes between Spanish-speaking outpatient groups randomly assigned to honest and feigning conditions. The psychometric properties of the Spanish SIRS primary scales were evaluated and the reliability and discriminant validity of the Spanish SIRS scales was also examined. Additionally, the study examined the participants' level of identification with American culture to see if significant differences existed between response styles on the Spanish SIRS and levels of acculturation.

Research Questions and Hypotheses

Hypothesis

1. Participants in the "feigning" condition will achieve higher scores on each primary scale of the Spanish SIRS than participants in the honest condition.

This hypothesis examines whether there are significant differences in performance between both experimental groups for each primary scale of the Spanish SIRS and explores the corresponding effect sizes. It is expected that, consistent with past SIRS research, the feigning group will generally achieve scores which are much higher than the control group (i.e., the "honest" condition).

Research Questions

1. Will the primary scales of the Spanish SIRS evidence high scale homogeneity?
2. What is the interrater reliability of the Spanish SIRS?
3. How accurate is the SIRS classification of feigning and honest conditions when applied to the Spanish SIRS?
4. Will participants with different levels of acculturation within the honest and feigning conditions have significantly different scale elevations on the Spanish SIRS?

The final research question explores whether there are significant differences in Spanish SIRS scores for people who have different levels of identification with American culture, based on scores from the ARSMA-II. Of particular interest are each participant's scores on the Anglo Orientation Subscale (AOS) and Mexican Orientation Subscale (AOS); individuals who score highly on AOS identify more strongly with American cultural values, whereas individuals who score highly on MOS identify more strongly with the cultural values of their ethnic group.

Supplementary Question

1. Will participants with different diagnostic constellations, as determined by the information obtained by the MINI, have significantly different scale elevations on the primary scales of the Spanish SIRS?

CHAPTER 2

METHOD

Study Design

The current study used a between-subjects simulation design involving two conditions ("honest" and "feigning") with outpatients voluntarily receiving mental health treatment from a staff of psychologists, psychiatrists, and social workers. To improve internal validity, participants were randomly assigned to standard and experimental conditions with manipulation checks to ensure adherence to these conditions. In addition, simulators were provided monetary incentives for convincing presentations. The original SIRS validation utilized the same experimental design to validate the measure (Rogers et al., 1991). To improve external validity, the study was conducted in an outpatient setting, Centro de Mi Salud, designed specifically for Hispanic American patients with most mental health services provided in Spanish.

Participants

The initial sample of 90 Spanish-speaking Hispanic outpatients, aged 18 years and older, was recruited from Centro de Mi Salud, an outpatient mental health center in Dallas, Texas. Centro de Mi Salud specializes in providing low-cost mental health services to people of low socioeconomic status whose primary language is Spanish. Inclusion criteria were minimal to maintain the representativeness of the sample: age (at least 18) and Spanish as the primary language. Given the nature of the setting, all adults met the language criterion. For those giving informed consent, the only exclusion criterion was severe psychotic symptoms that impaired the individual's ability to understand and respond relevantly to the Spanish SIRS. However, no participants were excluded based on this criterion.

Materials

Demographics Questionnaire

This brief interview-based questionnaire asked participants to report their age, occupation, gender, and ethnicity/race, as well as their reason for seeking treatment at the community mental health center and any current psychological diagnoses (see Appendix A for a copy of the demographics questionnaire).

Acculturation Rating Scales for Mexican Americans—2nd Edition (ARSMA-II)

The ARSMA-II (Cuellar, Arnold, & Maldonado, 1995) is among the most widely used and researched acculturation scales (Gamst et al., 2002). It contains two subscales with good internal consistency: the Anglo Orientation Subscale (AOS; Cronbach's alpha = .86) and the Mexican Orientation Subscale (MOS; Cronbach's alpha = .88), which are combined to produce a rating describing a person's degree of acculturation. One important advantage of the ARSMA-II is that its Spanish language version has been researched and validated for use with Spanish-speaking populations, unlike translations of other acculturation measures, whose psychometric properties have yet to be determined for Spanish translations (Malcarne, Chavira, Fernandez, & Liu, 2006).

Spanish SIRS

A Spanish translation of the Structured Interview of Reported Symptoms (SIRS) was administered to each participant. The SIRS is a structured interview designed to assess the feigning of mental disorders and related response styles. The Spanish translation of the SIRS uses a multiple-interviewer backtranslation procedure. The English version of the SIRS has good

to excellent internal reliability with alpha coefficients ranging from .77 to .92 for the primary scales with adequate to good estimates of .66 to .82 for the supplementary scales. Interrater reliability for its scales ranges from .95 to 1.00 (Rogers, 2001). Based on the primary scales of the SIRS, individuals can be classified as feigning, nonfeigning, and indeterminate. Research by Rogers, Gillis, Bagby, and Monteiro (1991) shows that the SIRS consistently detects large differences between those responding honestly and those feigning, thereby providing strong evidence for construct validity.

MINI – Spanish Version

The Mini International Neuropsychiatric Interview (MINI; Sheehan et al., 1998) was administered to participants to assess symptoms of Axis I disorders. The MINI is a structured interview which requires the examinee to respond "yes" or "no" to questions about specific symptoms of common disorders. The MINI was originally validated in a multi-national study involving more than 600 patients with mental disorders. It demonstrated excellent interrater reliability in both English (Sheehan et al., 1998) and French (Lecrubier, Sheehan, Weiller, Amorim, Bonora et al., 1997). Its reliability has been maintained with other translations (Japanese translation: Otsubo, Tanaka, & Koda, 2005; Italian version: (Rossi, Alberio, & Porta, 2004), but has not formally evaluated for the Spanish translation.

Procedure

The study received ethical approval from the Institutional Review Board (IRB) at the University of North Texas and administrative approval from Centro de Mi Salud before data collection began. All participants were provided informed consent in Spanish for their

involvement in the study. The procedure began with providing potential participants with written consent forms, which were also read aloud by a researcher. By adopting this procedure, issues of limited literacy were addressed without any perceived stigmatization.

The study utilized three advanced doctoral students in clinical psychology, who received specialized training in structured interviews, including Axis I interviews and the SIRS via coursework and research training. They were bilingual with a high level of fluency in Spanish. Prior to their involvement in the study, they reviewed and practiced the Spanish SIRS administration.

Phase I

Following the written informed consent in Spanish, the first researcher asked participants all questions on the demographics questionnaire and recorded their answers. All participants were then evaluated for their level of acculturation on the ARSMA-II, a straight-forward measure of their activities and cultural preferences. The ARSMA-II was placed first to help facilitate rapport; it was followed by the MINI that screens for symptoms of common Axis I disorders.

After Phase I was completed, the first researcher introduced each participant to the Phase II condition for the Spanish SIRS, either the feigning or honest condition. The instructions were sealed in a white envelope and placed into each testing packet in a quasi random fashion. Neither the investigator nor the participant knew the experimental condition until the envelope was opened just prior to explaining the instructions. At that point, the investigator remained masked to the condition. As part of these instructions, each participant was told they would earn

$10.00 as an incentive for successfully completing their assigned task. After the instructions were explained, participants had an opportunity to ask questions before beginning Phase II.

Phase II

A second researcher, masked to both the results of Phase I and the participant's condition (honest or feigning), then administered the Spanish SIRS. After the completion of the Spanish SIRS, the manipulation check was performed. For the manipulation check, participants were asked to recall the experimental instructions in their own words and rate their overall effort. Of the 90 participants interviewed, four could not recall or did not follow the instructions, and an additional four reported doing a "poor" job. Both groups were excluded from any further analyses. An additional two participants dropped out before the completion of the study, stating time conflicts. Thus, the final sample was composed of 80 participants.

After the debriefing, all participants were thanked and paid the $10.00 incentive. To avoid any ethical concerns (e.g., penalizing more impaired participants) participants were paid regardless of whether they declared they properly followed experimental instructions. As noted, the basic design required two bilingual researchers for each phase. In eight cases, a third bilingual researcher was available to examine interrater reliability with both researchers independently scoring the same Spanish SIRS administration.

Experimental Conditions

Using a quasi-random procedure, participants were assigned to one of two conditions in a between-subjects design. A white envelope containing a set of written instructions and the participant's experimental condition was placed in each administration packet. The same type of

white envelope was used for both conditions, so that the experimenter would be masked to the experimental condition of any particular participant. The researcher simply drew any administration packet from the bag of testing materials and remained unaware of which instructions it contained until it was time to give instructions to the participant.

It was crucial that the feigning condition be relevant to participants, engaging, and easily understood (Rogers & Cruise, 1998). For this reason, they were presented with a scenario describing their task. Because of their expected familiarity with disability benefits and the mental healthcare system, participants in the feigning condition were asked to simulate persons who would qualify for full disability insurance because of their psychological impairment. Specifically, they were told to fabricate psychological symptoms and associated features during their interview. They were asked to pretend the study was a disability evaluation, warned to be convincing in their presentations, and challenged to "fool the examiner" into believing they were responding truthfully in their portrayal of feigned symptomatology (see Appendix B for scenario and instructions).

As this is the initial study of the Spanish SIRS, participants in the feigning condition were not coached by giving them additional knowledge of symptom presentations or how to malinger. This decision was made for two primary reasons. First, it is likely that these outpatients have more insight into psychological disorders than the general population because of their own histories of mental health treatment. Additionally, it was crucial to avoid influencing how this cultural group might respond to questions on the Spanish SIRS.

For the honest condition, participants were asked to be truthful and forthcoming about their current symptoms. They were not presented with a scenario. Instead, their instructions stressed the importance of this research in creating a valid test that would be of optimum use in

helping Hispanic American individuals undergoing psychological evaluations (see Appendix B for complete instructions).

After completing all measures, the second researcher conducted a manipulation check with each participant (see Appendix C). At this time, volunteers were asked to recall the experimental instructions in their own words as the researcher recorded their responses. All participants were debriefed and informed about the general goals of the study. They were, then, given the $10.00 incentive previously described at the beginning of the study.

As a critical last step that would later determine whether a participant's data should be excluded from analysis, participants were asked to rate how well they completed their assigned task ("poor," "good," "excellent"). Participants also evaluated the effort they put forth in following their instructions on a scale from 0 to 100%. Participants who could not remember their experimental instructions or report "poor" effort were excluded from data analysis so that their failure to properly complete the task did not artificially alter SIRS scale elevations or classificatory accuracy for the Spanish SIRS. The specific questions posed to participants can be found in Appendix C.

For eight participants, a second Spanish-speaking researcher observed the Spanish SIRS administration. This second researcher independently scored the Spanish SIRS responses, in order to establish interrater reliability.

CHAPTER 3

RESULTS

The final sample was composed of Hispanic American outpatients whose country of origin was predominantly Mexico, with smaller representations from other countries. The cultural breakdown is presented in Table 3.

Table 3

Percentage of Sample from Each Represented Country of Origin

Country of Origin	Number of Participants	Percentage
Mexico	75	94.9%
El Salvador	2	2.5%
Guatemala	1	1.3%
Honduras	1	1.3%

Because of Dallas' proximity to Mexico, it is predictable that the vast majority of participants (94.9%) originated from that country. There were small representations from three other countries totaling 5.2% of the sample. Table 4 explores participants' overall level of identification with Hispanic culture.

Table 4

Acculturation Level for Spanish-speaking Outpatients

Level of Acculturation	Number of Participants	Percentage
Traditional	61	76.3
Marginalized	8	10.0
Bicultural	7	8.8
Assimilated	2	2.5

Most participants (61 or 76.3%) identified predominantly with Mexican culture and were classified with a traditional perspective. With two (2.5%) unclassified by the ARSMA-II, most of the remaining were grouped as marginalized (8 or 10.0%) or bicultural (7 or 8.8%). As expected, very few (2 or 2.5%) were classified as assimilated.

Table 5

Age and Gender Composition of the Sample

Gender	Male (n = 26)		Female (n = 54)		Total (n = 80)	
Age	*M*	*SD*	*M*	*SD*	*M*	*SD*
	31.4	10.8	35.9	11.3	34.45	11.3

The sample was composed of 54 women and 26 men, with participant age ranging from 18 years to 70 years (M = 34.5). Additional demographic information unique to this sample is presented in Tables 6, 7, and 8.

Table 6

Participants' Level of Education

Level of Education	Number of Participants	Percentage
Elementary School	18	22.5%
Middle	21	26.3%
High School	26	32.5%
GED	1	1.3%
Vocational School	4	5.0%
Some College	4	5.1%
Bachelor's Degree	5	6.3%
Advanced Degree	0	0.0%

A substantial 48.8% of the sample had less than a high school education while only 11.4% received any higher education. Exact data on where participants were educated is

unavailable. However, most of the sample was likely educated outside of the United States, as an overwhelming 88.5% of the sample reported being first-generation U.S. immigrants. 10.3 % of participants were second-generation, and one participant (1.3% of the sample) described herself as fifth-generation. Lower levels of education, as well as education received outside of the U. S. could have an impact on responses to certain Spanish SIRS items. These issues are further addressed in the Discussion.

Table 7

Employment Status of Sample Participants

Socioeconomic Status	Number of Participants	Percentage
Unemployed	41	52.6%
Part-Time	13	16.7%
Full-Time	19	24.4%
Disabled	5	6.4%

Centro De Mi Salud specializes in providing low-cost mental health services to those with financial need. Thus, it is not surprising that 59% of the sample reported being unemployed or receiving disability insurance.

Table 8

Reported Socioeconomic Status of Sample Participants

Socioeconomic Status	Number of Participants	Percentage
Lower	50	64.9%
Middle	26	33.8%
Upper	1	1.3%

With the majority of the sample reporting unemployment or disability (59%) and less than a high school education (48.8%), it is understandable that a majority (64.9%) also reported

being of low socioeconomic status. The sample had only one participant (1.3% of the sample) reported being a member of the "Upper" class.

A variety of Axis I disorders were represented by the sample of outpatients in this study. Based on the MINI, the following diagnostic categories were represented among the outpatient participants: prominent psychotic symptoms (21.3%), major depression (37.5%), other mood disorders (10%), anxiety disorders (33.8%), and substance abuse disorders (13.8%). Analyses regarding possible effects that symptom constellations might have on scale elevations of the Spanish SIRS are addressed in a later section of the chapter.

Reliability of the Spanish SIRS

The first research question addressed the internal consistency of Spanish SIRS primary scales. Alpha coefficients and inter-item correlations were calculated (see Table 9).

Table 9

Internal Consistencies, Interrater Reliabilities, and Standard Errors of Measurements (SEM) for the Spanish SIRS Primary Scales

SIRS	Alpha	Mean Inter-Item Correlations	Interrater *r*	SEM (α)
RS	.81	.35	1.00	1.12
SC	.89	.44	1.00	.98
IA	.84	.43	1.00	.66
BL	.96	.59	1.00	1.02
SU	.95	.52	.99	2.00
SEL[a]	NA	NA	1.00	NA
SEV[a]	NA	NA	.99	NA
RO	.76	.23	.98	.58
Average	.89	.43	.995	1.06

Note. Because of their deliberate distortions, feigners are not expected to produce uniform results; therefore, SEMs are calculated using the *SD*s under the honest condition. [a] SEL and SEV involve counts across several detection strategies; thus, their unidimensionality is not assumed and α is not calculated.

For scales to be comparable to the English-language version of the SIRS, the objective is to achieve high alphas (i.e., $\geq$ .80) for each Spanish SIRS scale. The Alpha coefficients of most applicable primary scales were greater than .80 with a mean of .89. It indicates that items within each scale measure the same general construct. The sole exception was RO (alpha = .76) which nearly met the criterion. This could be because RO is the only SIRS scale where evaluators must combine their own clinical observations with the participants' self-report, rather than relying solely on the participants' self-report. Overall, mean inter-item correlations are moderate. Again, RO is the only scale that differs, exhibiting a weak inter-item correlation.

The reliability of individual test scores, as expressed by the standard error of measurement (SEM), is an integral component of reliability (Anastasi, 1988). The SEM (α) is provided rather than SEM (*r*) because of the small number of reliability cases. Using this approach, most SEMs (see Table 9) are about 1 point (*M* = 1.06; range from .58 to 2.00). In contrast, the SU scale which has much more variability (*SD* = 8.96, SEM = 2.00) among patients in the honest condition than the other SIRS primary scales.

Research Question 2 investigates the interrater reliability of Spanish SIRS scales. For 8 cases, a second researcher also independently scored each Spanish SIRS administration. Overall, the interrater reliabilities are very high (see Table 9), ranging from .98 to 1.00 (*M* = .995. Such high numbers are expected for a fully structured interview, such as the Spanish SIRS.

Research Question 3 sought to investigate the accuracy of SIRS cut scores for distinguishing simulators from those in the honest condition. Utility estimates using the original SIRS classification criteria were used to determine the sensitivity, specificity, positive predictive power (PPP), and negative predictive power (NPP) of the Spanish SIRS in the assessment of malingering. According to the original scoring rules of the SIRS, the basic classification of

feigning involves (a) one or more primary scales in the definite feigning range, or (b) three or more primary scales in the probable feigning range. For marginal cases (e.g., one or two scales in the probable feigning range), the SIRS total score > 76 can be employed. When these rules were applied to the Spanish SIRS in this sample, the overall classification rate was high at .88. Sensitivity (.90) and specificity (.85) were well balanced. Similar estimates were found for positive predictive power (PPP = .86) and negative predictive power (NPP = .89). The false-positive rate was 15%.

Accuracy of the Spanish SIRS

The discriminability of SIRS primary scales are of critical importance to their clinical usefulness. Hypothesis 1 predicted that individuals in the feigning condition would produce higher Spanish SIRS primary scale scores than those in the honest condition. As shown in Table 10, one-way analyses of variance (ANOVAs) conducted on each of the Spanish SIRS scales demonstrated large differences between the two groups.

Table 10

Differences on the Spanish SIRS Primary Scales between Honest and Feigned Presentations

	Honest (n = 40)		Feigned (n = 40)			
SIRS scales	*M*	*SD*	*M*	*SD*	*F*	*d*
RS	1.92	2.57	8.73	4.27	73.07	1.92
SC	2.03	2.94	11.33	5.61	84.52	2.07
IA	0.85	1.65	6.75	4.22	66.53	1.84
BL	4.10	5.11	21.56	8.58	122.00	2.47
SU	9.25	8.96	25.48	8.42	69.71	1.87
SEL	8.03	6.93	24.45	7.63	101.54	2.25
SEV	5.33	6.65	22.58	9.01	94.96	2.18
RO	0.62	1.18	3.73	2.94	37.78	1.38

Note. For all *F* ratios, $p < .0001$

Using Rogers' (2008) guidelines (i.e., "large" effect size, $d \geq 1.25$; "very large," $d \geq 1.50$), Spanish SIRS primary scales generally produced very large effect sizes (*M d* = 2.00; range from 1.38 to 2.47) between feigned and genuine conditions. Interestingly, SIRS scales using amplified detection strategies (i.e., BL, SU, SEL, and SEV) produced somewhat higher effect sizes (*M d* = 2.19 versus *M d* = 1.80) than those utilizing unlikely detection strategies (RS, SC, IA, and RO) for this population. Overall, these results provide strong evidence that the Spanish SIRS primary scales clearly differentiate between feigned and genuine conditions among Spanish-speaking Hispanic outpatients.

Table 11

Differences on the Spanish SIRS Supplementary Scales between Genuine and Feigned Presentations

	Genuine (n = 40)		Feigned (n = 40)			
SIRS scales	*M*	*SD*	*M*	*SD*	*F*	*d*
DA	2.79	2.51	9.05	4.97	49.40	1.58*
DS	18.41	9.39	29.10	8.91	26.94	1.17*
OS	2.10	2.06	6.65	4.40	39.47	1.32*
SO	2.12	1.47	3.60	0.93	28.40	1.20*
INC	3.40	3.51	4.65	4.73	1.80	0.30

* Denotes $p < .0001$ for *F* ratios

Although the RO scale was the lowest effect size, it is still considered large (*d* = 1.38). Of note, RO is the only scale where clinicians are asked to use their judgment based on clinical observation. High scores on RO are reserved only for responses where the clinician has clear evidence of inconsistency between their observation and the client's self report (Rogers, Bagby, & Dickens, 1992). A limitation of this study is that the researcher administering the Spanish SIRS only observed each participant during the SIRS administration, as they needed to remain masked to participant performance during the Phase I of testing. Restrictions on the ability to

observe each outpatient might have led to RO scores being slightly lower than other primary scales.

The feigning group attained significantly higher scores than the honest group on every supplementary scale of the Spanish SIRS except INC. Of these, OS (d = 1.32) and DA (d = 1.58) produced a very large effect sizes. For the remaining, DS and SO produced moderate effect sizes, and INC demonstrated no significant difference between feigners and honest responders.

Acculturation and the Spanish SIRS

The effects of acculturation on the Spanish SIRS primary scales was investigated in order to determine the generalizability of the Spanish SIRS across Spanish speaking individuals who differ in their cultural identification (Anastasi, 1988; Okazaki & Sue, 1995; Wagner & Gartner, 1997). Research Question 4 sought to test the effects of acculturation on Spanish SIRS primary scales. It was initially proposed that participants from the control condition would be divided into groups that reflect their level of acculturation, based on the scores they attained on the ARSMA-II. However, once data collection was completed, it became apparent that there would not be sufficient statistical power to conduct these analyses for each of the four acculturation groups, as the vast majority (76.3%) of participants fell into the Traditional group. For this reason, it was decided to analyze correlations between two acculturation groups (Traditional vs. Non-Traditional), rather than the four groups, as originally planned. These analyses were conducted for participants in the honest condition and the feigning condition separately, in order to provide initial data on whether acculturation may facilitate feigning.

As seen in Table 12, nearly all correlations are non-significant. The sole exception is that level of acculturation has a moderate relationship with scores on the IA scale for participants in

the honest condition (r = -0.44). The absence of statistically significant correlations for the remainder of the Spanish SIRS primary scales among both honest and feigning participants suggests that, for the most part, level of acculturation is not significantly correlated with performance on the Spanish SIRS.

Table 12

Correlations of Primary Scale Scores and Acculturation (Traditional vs. Non-Traditional) for the Spanish SIRS

Spanish SIRS Scales	Honest		Feigning	
	r	*p*	*r*	*p*
RS	-0.11	0.52	.01	.95
SC	-0.21	0.22	.09	.59
IA	-0.44	0.01	-.05	.78
BL	-0.10	0.55	-.00	.99
SU	-0.05	0.77	-.02	.89
SEL	-0.06	0.71	-.03	.86
SEV	-0.07	0.64	.00	.997
RO	-0.001	0.99	-.11	.52

Table 13 illustrates means, standard deviations, and effect sizes for both experimental conditions based on level of acculturation. All effect sizes are large to very large (d ranges from 1.45 to 2.39 among traditionally oriented participants and d ranges from 1.29 to 2.49 among "other" levels of acculturation). Despite a moderate correlation between scale score and level of acculturation, the IA scale produced very large effect sizes when distinguishing between feigners and honest responders for participants with both Traditional (d = 1.91) and Other (d = 1.50) cultural identifications. In summary, the Spanish SIRS still effectively distinguishes between feigners and honest responders for both Traditional and Non-Traditional (Other) levels of acculturation.

Table 13

Differences on the Spanish SIRS Primary Scales between Genuine and Feigned Presentations for level of Acculturation (Traditional vs. Other)

	Honest				Feigning				Effect Sizes	
	Traditional		Other		Traditional		Other		Trad.	Other
SIRS	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*d*
RS	1.62	2.14	2.22	3.19	8.53	4.51	8.40	3.51	1.93	1.87
SC	1.55	2.13	2.89	4.37	11.31	5.69	9.80	6.87	2.23	1.29
IA	0.38	0.78	2.00	2.69	6.41	4.30	7.00	4.36	1.91	1.50
BL	3.76	4.90	4.90	6.08	21.34	9.00	21.40	7.73	2.39	2.49
SU	9.00	8.87	10.00	10.09	25.00	8.68	25.60	9.32	1.82	1.58
SEL	7.72	6.73	8.70	8.14	24.13	8.15	24.80	6.42	2.19	2.10
SEV	5.03	6.58	6.20	7.48	22.22	9.26	22.20	10.33	2.12	1.89
RO	0.55	1.21	0.56	0.88	3.66	2.72	4.60	4.56	1.45	1.48

Note. These effect sizes are between honest and feigning participants.

I analyzed effect sizes of each primary scale for the Traditional group in both the honest and feigning conditions. We then compared their effect sizes, with those of the total Spanish-speaking sample, and the original validation sample for the English language version of the SIRS (Rogers, Bagby, & Dickens, 1992) to see if there were substantial differences. In computing these effect sizes, we found that the values were quite similar between the Traditional group and the total sample. These similarities are expected, given the overlap in participants between the two groups (61 of 80 or 76.3% was classified as Traditional). Nonetheless, there were very large effect sizes found in those individuals identifying with a non-American (i.e., Mexican) culture. In most cases, the effect sizes are larger than in the original validation sample of the SIRS; RO is the only exception.

Table 14

Differences between Genuine and Feigned Presentations for Traditional Hispanic Outpatients with Comparisons of Effect Sizes for the Total Sample (Traditional and Other combined) and the Original English Validation

	Traditional Hispanic Outpatients							
	Honest (n =29)		Feigning (n =32)				Total Sample (n = 80)	Original Validation Sample (n = 270)
	M	*SD*	*M*	*SD*	*F*	*d*	*d*	*d*
RS	1.62	2.14	8.53	4.51	56.49	1.93	1.92	1.83
SC	1.55	2.13	11.31	5.69	75.54	2.23	2.07	1.48
IA	0.38	0.78	6.41	4.30	55.21	1.91	1.84	1.20
BL	3.76	4.90	21.34	9.00	87.16	2.39	2.47	1.87
SU	9.00	8.87	25.00	8.68	50.63	1.82	1.87	1.79
SEL	7.72	6.73	24.13	8.15	72.61	2.19	2.25	1.98
SEV	5.03	6.58	22.22	9.26	68.55	2.12	2.18	1.95
RO	0.55	1.21	3.66	2.72	32.00	1.45	1.38	1.78
M d						2.00	2.00	1.74

Supplementary Analyses

A supplementary question in this study seeks to explore whether participants with different diagnostic presentations exhibit different scale elevations on the primary scales of the Spanish SIRS. Three main symptom constellations are represented in the sample; these diagnostic categories, as determined by the MINI are: Psychotic symptoms, Major Depression, and Generalized Anxiety Disorder. The first step in addressing the supplementary question was to examine the relationships between symptoms of each constellation and Spanish SIRS scale scores for each of the three groups listed above via correlations. Table 15 illustrates the strength of correlations among participants with psychotic symptoms in both the honest and feigning conditions.

Table 15

Correlations of Primary Scale Scores for the Spanish SIRS and Psychotic Symptoms on the MINI

Spanish SIRS scales	Honest *r*	Feigning *r*
RS	0.49**	0.20
SC	0.63**	0.17
IA	0.28	0.33*
BL	0.29	0.19
SU	0.08	0.14
SEL	0.12	0.18
SEV	0.20	0.16
RO	0.10	-0.08

* Denotes significance at the 0.05 level (2-tailed); ** Denotes significance at the 0.01 level (2-tailed)

Significant correlations were found between psychotic symptoms and scale scores for RS, SC, and IA, with the strongest relationships found for individuals in the honest condition. These three scales utilize unlikely detection strategies and are comprised of questions that often reflect psychotic content. Thus, it follows that patients diagnosed with psychotic disorders would attain higher scores on these scales. Group size for participants with psychotic disorders in the honest condition was not large enough ($n = 4$) to conduct a Chi-square test for Independence to determine whether there was a relationship between psychosis and responses on the most commonly endorsed items on the RS and SC scales. There is a weak correlation between psychotic symptoms and IA scores for participants in the feigning condition only. Possible effects of these correlations on the effectiveness of the Spanish SIRS in distinguishing feigners from honest responders are addressed later (see Tables 18 and 19). First, relationships between scale scores and two additional diagnostic constellations are discussed.

Table 16 shows the strength of correlations between scale scores and symptoms of Major Depression for participants in both the honest and feigning conditions. There were weak to

moderate correlations (*r* ranging from 0.37 to 0.42) for each of the scales that utilize amplified detection strategies (BL, SU, SEL, and SEV) for participants in the honest condition.

Table 16

Correlations of Primary Scale Scores for the Spanish SIRS and Symptoms of Major Depression on the MINI

Spanish SIRS scales	Honest *r*	Feigning *r*
RS	0.25	0.12
SC	0.34*	0.15
IA	0.32*	0.29
BL	0.37*	0.14
SU	0.40**	0.14
SEL	0.42**	0.18
SEV	0.39*	0.11
RO	0.16	0.10

* Denotes significance at the 0.05 level (2-tailed); ** Denotes significance at the 0.01 level (2-tailed)

Weak correlations also existed for two scales using unlikely detection strategies (SC and IA). There is some overlap in the diagnostic constellations, as psychotic symptoms tended to be comorbid with other Axis I disorders. It could be that the large number of significant correlations here is a product of that. Possible effects of these correlations on the effectiveness of the Spanish SIRS in distinguishing feigners from honest responders are addressed in the Discussion.

Table 17 explores the relationships between scale scores and symptoms of Generalized Anxiety Disorder (GAD). Weak correlations were found between symptoms of GAD and scale scores for SU, SEL, and SEV for individuals in the honest condition (r = .40, .37, and .34, respectively). These three scales utilize amplified detection strategies. High scores indicate high endorsement of symptoms and their severity. It can be expected that patients diagnosed with mood disorders would attain higher scores on these scales, particularly when they report their genuine symptoms in the honest condition. Possible effects of these correlations on the

effectiveness of the Spanish SIRS in distinguishing feigners from honest responders are addressed in Tables 18, 19, and 20.

Table 17

Correlations of Primary Scale Scores for the Spanish SIRS and Symptoms of Generalized Anxiety Disorder on the MINI

Spanish SIRS scales	Honest *r*	Feigning *r*
RS	0.01	0.17
SC	0.13	0.02
IA	0.07	0.13
BL	0.23	0.21
SU	0.40**	0.07
SEL	0.37*	0.16
SEV	0.34*	0.13
RO	0.31	0.16

* Denotes significance at the 0.05 level (2-tailed); ** Denotes significance at the 0.01 level (2-tailed)

For each of the diagnostic groups investigated, results show that false-alarm rates are appreciably lower for participants lacking genuine symptoms of a psychological disorder (see Table 18). Participants classified with psychotic symptoms exhibited the highest false-alarm rate (40%), while those without Major Depressive Disorder (MDD) had the lowest false-alarm rate (5.0%). As it turns out, there was a high rate of comorbidity among Axis I disorders for this sample. High overlap among patients with psychotic symptoms, MDD, and GAD, suggests the comparisons in Table 18 are far from clean, with various individuals being included in more than one group simply because they met criteria for more than one diagnosis. In an attempt to lessen the overlap when comparing groups, Table 19 looks at false-alarm rates between individuals with with *only* MDD, and those with *only* GAD.

Table 18

Differences in Spanish SIRS False-alarm Rates Among Patients with Symptoms of Possible Comorbid Disorders Including Psychosis, Major Depression, and Generalized Anxiety Disorder

	Comorbid with Psychosis		Comorbid with Major Depression		Comorbid with GAD	
	Present	Absent	Present	Absent	Present	Absent
Number of False-positives	2	4	5	1	4	2
False-alarm Rate	40.0%	11.0%	29.0%	5.0%	25.0%	15.0%
Total *N*	17	63	30	49	27	53

Note. The MINI provides screening information and not diagnoses per se; these categories are considered "possible disorders" because the MINI Spanish version has not been validated.

Table 19

Differences in Spanish SIRS False-alarm Rates Among Patients with Possible Disorders

	Major Depression Only		GAD Only	
	Present	Absent	Present	Absent
False-alarm Rate	33.0%	5.0%	17.0%	15.0%
Number of False Positives	2	1	1	2
Total N	12	49	9	53

The small number of psychotic individuals in the honest condition does not allow for an analysis of psychotic individuals with no other Axis I disorders. Only a total of four participants fall into this category. Small group size is a limitation of this study, making it nearly impossible to adequately study group trends for the supplementary question regarding diagnostic categories. The false-alarm rate among individuals with MDD only (no psychotic symptoms and no GAD) is slightly higher than the rate among individuals with MDD and other comorbid Axis I disorders (see Table 18). However, group size becomes highly discrepant. For example, Table 19 compares an *n* of 12 to an *n* of 49. The false-alarm rate of the Spanish SIRS decreases from 25.0% to 17.0% when looking at participants with GAD only but, small group size compromises the power of these analyses and the conclusions that can be drawn from them.

Differences between honest responders and feigning participants are explored via ANOVAs and Cohen's *d* for these three diagnostic groups in Table 20.

Table 20

Effect Sizes between Genuine and Feigned Presentations for Symptoms of Psychosis, Major Depression, and Generalized Anxiety Disorder

SIRS	Psychosis		Major depression		GAD		Total
	Present	Absent	Present	Absent	Present	Absent	sample
RS	1.29	1.99	1.86	1.96	2.48	1.92	1.92
SC	1.37	2.18	2.08	2.10	2.70	2.09	2.07
IA	1.91	1.75	2.43	1.70	2.70	1.61	1.84
BL	1.66	2.67	2.56	2.54	3.64	2.48	2.47
SU	1.49	1.92	1.77	2.36	1.81	2.44	1.87
SEL	1.72	2.35	2.15	2.57	2.59	2.62	2.25
SEV	1.47	2.35	2.05	2.43	2.84	2.44	2.18
RO	1.49	1.34	1.43	1.30	1.28	1.50	1.38
M	1.55	2.07	2.04	2.12	2.51	2.14	2.00

Note. The MINI provides screening information and not diagnoses per se; these categories are considered "possible disorders" because the MINI Spanish version has not been validated.

Overall, the Spanish SIRS effectively distinguished between feigners and honest responders regardless of whether symptoms of a disorder were present, with all scales producing large to very large effect sizes (*d* ranges from 1.28 to 3.64). Looking at specific symptom constellations, effect sizes tend to be larger in the absence of psychotic symptoms except for the IA and RO scales.

Much more variability was noted in trends among effect sizes for the remaining symptom constellations. For Major Depression, effect sizes also tend to be larger in the absence of symptoms. Notable exceptions include the IA, BL, and RO scale, which follow the opposite trend. This opposite trend seems to be much more common for symptoms of GAD, where effect sizes tend to be larger when symptoms are present. Exceptions to this trend for GAD are the SU, SEL, and RO scales.

CHAPTER 4

DISCUSSION

Ethical guidelines from the American Psychological Association require that psychologists working with ethnically, linguistically, and culturally diverse populations should recognize these characteristics as important factors affecting a person's experiences, attitudes, and psychological presentation (Bersoff, 2004). Psychologists can easily conclude that these varying factors can also have important effects on assessment results when evaluated by standardized testing measures. This issue of diversity in assessment is especially important when considering an individual's preferred language and when using test translations, as a translated measure does not necessarily retain the psychometric properties of the original language version (APA, 1993).

Throughout recent years, different professional organizations have addressed issues of diversity and created guidelines and standards for addressing these issues within the realm of psychological testing. For example, the Standards for Educational and Psychological Testing from the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME, 1999) addressed language and diversity by specifying that any oral or written test also measures an examinee's verbal skills. According to the Standards, the reliance on verbal abilities creates a particular concern for individuals whose primary language is not the original language of the test. These standards conclude that "in such instances, test results may not reflect accurately the qualities and competencies intended to be measured" (AERA, APA, NCME, 1999, p. 91). On this point, translated tests can create test bias, the possibility for misdiagnosis, and the serious misinterpretation of test results (Dana, 1993).

Issues of test bias are magnified when translated versions of assessment measures are used in professional settings. The Test Translation and Adaptation Guidelines developed by the International Test Commission (ITC; Hambleton, 2001) specify that test developers must apply appropriate research methods and statistical techniques to establish the validity of each translated test for the new target population. Only tests that have been formally translated and subsequently validated as translated tests should be used in clinical practice (Hambleton, 2001). Currently, there are few tests that have been adequately validated for use with Hispanic American populations whose primary language is Spanish. Because of the pressing need for Spanish measures, the current study focused on reliability, validity, and other psychometric properties of the Spanish SIRS.

Before discussing results specific to the Spanish SIRS, an overview regarding the current availability of Spanish language assessment measures is explored below with an emphasis on their clinical utility with monolingual Hispanic Americans. The paucity of well-researched Spanish language testing measures is clearly evident in many domains of psychological assessment, such as intellectual functioning, psychopathology, and specialized issues which include, but are not limited to, response styles. What follows is an overview of current assessment practices used with Hispanic American clients for various domains of psychological testing.

Intelligence and Cognitive Testing for Hispanic American Clients

The scarcity of normative data pertaining to minority populations is apparent in many psychological assessment measures. The current study was designed after reviewing the state of the art for Spanish language translated measures. What follows is a specific look at the

provisions presently made by test developers to address the needs of Spanish-speaking individuals on commonly used testing instruments. These advances, used in conjunction with the ITC Guidelines (Hambleton, 2001) form the basis for the current study on the Spanish SIRS.

In the realm of intelligence and cognitive testing, researchers have long since pointed out that demographic variables such as age, gender, and culture can significantly affect an individual's performance on cognitive tests. When a person's demographic characteristics are different than the normative group, the potential for misdiagnosis and inaccurate interpretation of their test scores increases. Heaton, Taylor, and Manly (2003) used all available data from the English language WAIS-WMS co-norming project to provide new cut scores that correct for demographic influences on an individual's test scores. Results of this endeavor consistently found small but significant effect sizes for English-speaking Hispanic American individuals. Hispanic Americans generally achieved lower scores than their European American counterparts when both groups were tested in English. Using standard norms, a substantial 15 to 25 percent of Hispanic individuals were misclassified as "impaired" on each factor of the WAIS-III when corrections were already made for other factors such as age, gender, and level of education. In order to reduce ethnicity bias, normative adjustments were suggested by Heaton, et al. 2003). Fortunately, when using the resulting corrected norms, individuals have nearly the same likelihood (approximately 15%) of being misclassified as their European American counterparts.

When examining WAIS-III performance within groups of Hispanic American individuals tested in English, research shows they typically achieve lower scores on verbal tests than on non-verbal tests. Kaufman and Lichtenberger (2006) hypothesized that lower scores are a reflection of (a) unfair language demands placed on individuals for whom English is a second language, and (b) the cultural content of some verbal test items. The same researchers recommend

administering only the non-verbal portions of IQ tests to individuals when English proficiency is called into question. However, this practice compromises a thorough assessment, as it prevents the evaluator from determining both the Verbal IQ (VIQ) and Full Scale IQ (FSIQ) of an examinee and relies solely on the Perfomance IQ (PIQ). In light of the misclassification rates noted for Hispanic individuals, even on PIQ (Heaton, et al., 2003), mental health professionals should be cautious in interpreting all assessment results and use alternate cut scores when appropriate. However, when English is not an examinee's preferred language, cut scores for fluent English speakers may not be appropriate and omitting VIQ and FSIQ may be an examiner's only recourse (Kaufman & Lichtenberger, 2006). Such concerns are the primary reason specialized research should be conducted for each translation in order to determine how to best apply a test to a population for whom it might not have been originally intended (Hambleton, 2001).

Fortunately, some Spanish language measures are available for intelligence testing. Thus, so long as the evaluator is fluent in Spanish, the client can be tested in that language and the assessment does not have to omit information on a client's verbal skills. The Kaufman Brief Intelligence Test Second Edition (K-BIT 2; Kaufman & Kaufman, 2004b) measures both the verbal and non-verbal abilities of individuals aged 4 to 90, and has satisfactory reliability and validity for the English language version. For individuals with low English proficiency, the K-BIT 2 provides Spanish translations of test items within the same test kit as the English version. It also provides scoring options for different Spanish language responses. While having this translation is seemingly beneficial to individuals whose primary language is not English, the Spanish-speaking individuals were omitted from the normative sample, making it difficult to interpret their results. The lack of normative data is longstanding (see Sattler, 2001). It was

observed with the original version of the K-BIT (Kaufman & Kaufman, 1990) and with the Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004a) which also provides Spanish translations of test items and responses. Due to a lack of validation research for the KABC's Spanish language verbal items, Kaufman and Kaufman (2004a) include a warning in the manual, stating that the test is "not intended to be administered in Spanish; except for the Nonverbal Scale, it is for use with children who are proficient in English" (Kaufman & Kaufman, 2004a, p.1). Furthermore, without research testing the equivalence of Spanish and English versions of the K-BIT 2 or the KABC-II, as recommended by the ITC, clinicians cannot make an informed decision as to whether a bilingual individual should be assessed in either language (Hambleton, 2001).

The other Spanish language IQ measures that are available also suffer from a lack of validation research with normative samples of Spanish-speaking individuals. For example, the Spanish language version of the WAIS-III, known as the Escala de Inteligencia de Weschler para Adultos – Tercera Edicion (EIWA-III; Weschler, 2008) is now commercially available in the United States. The EIWA-III includes the same subtests and constructs as the WAIS-III and is published by Pearson, the same company as the English-language WAIS-III. This measure was developed and tested in Puerto Rico to ensure that items were culturally appropriate for Puerto Rican individuals speaking Spanish. With this population, the EIWA-III demonstrates high internal consistency with mean alpha coefficients ranging from .73 to .92 and mean standard error ranging .94 to 1.56 for subtests across all age groups (Pons, Flores-Pabon, Matias-Carrelo, Rodriguez, Rosario-Hernandez, Rodrigues, Herrans, & Yang, 2008).

To date, however, there are no published studies on the validity or reliability of the EIWA-III with other Spanish speaking populations. Additionally, no research compares its

psychometric properties to the English-language WAIS-III. If the EIWA-III is used for persons outside of Puerto Rico, this lack of psychometric validation goes against both the ITC standards, and the Standards for Educational and Psychological Testing which require that psychologists and other professionals refrain from using a translated version until the reliability and validity of that new measure has been established for each population with which it is used (AERA, APA, NCME, 1999; Hambleton, 2001). The danger in administering tests that have not been validated is that clinicians interpret the results based on an *assumption* that the test continues to function in the intended manner Fantoni-Salvador, 1997). Such assumptions effectively force minority individuals into inappropriate interpretative categories, thereby creating a substantial possibility for misdiagnosis and misinterpretation of test results (Dana, 1993; Todd, 2005). Clinicians must provide caveats while interpreting assessment data and tailor treatment recommendations to different groups of minority clients (Correa, & Rogers, in press).

A small amount of validation research has been conducted on a different Spanish translation of the WAIS-III entitled the Spanish WAIS-III (TEA Ediciones, 2001), adapted and published in Spain. Research using a Spanish-speaking monolingual sample from Spain demonstrates that this version of the Spanish WAIS-III supports the same four-factor structure as the English WAIS-III (Garcia, Ruiz, & Abad, 2003). However, no comparison was carried out to determine the equivalency of the tests and its normative data has only been established using the Spanish-speaking sample from Spain. Using a Spanish-speaking sample of Hispanic Americans in Chicago, Renteria, Li, and Pliskin (2008) have conducted the only published validation study on the TEA edition of the Spanish WAIS-III in the United States. Results found adequate reliability and criterion validity for the TEA Spanish WAIS-III. When used with the Hispanic American sample, Spanish WAIS-III subtests had an average internal consistency reliability that

was similar to the averages for the sample from Spain (using the Spanish WAIS-III) as well as the North American English-speaking sample (using the English language WAIS-III).

Renteria, et al. (2008) also identified various areas of bias within the Spanish WAIS-III. For example, they recommend one subtest (Letter-Number Sequencing) that should be omitted because its inadequate alpha coefficients indicate poor construct validity. If this subtest is included in analysis, Renteria et al. cautioned that scoring should be more lenient because the structure of the Latin American alphabet makes this task more difficult in Spanish than in English. Lastly, Renteria et al. (2008) highlighted specific areas where test bias exists in favor of Spaniards, but lower scores are seen for Spanish-speaking individuals from other Latin American cultures.

In summary, several options are available for Spanish language cognitive assessments, each with its strengths and weaknesses. An attractive quality of the K-BIT 2 and KABC 2 is that both the Spanish and English versions are included in the same test booklets, eliminating the need for evaluators to purchase two separate testing kits. A considerable drawback, however, is the absence of validation data for their Spanish versions. The EIWA-III has published validation data for Puerto Rican populations, however, its effectiveness with US populations has yet to be tested. Of the three Spanish language measures available, the most researched measure might be the least accessible to mental health professionals in the United States. The Spanish WAIS-III, published in Europe, is the sole measure with validation data available for US populations and the only measure for which specific areas of potential test bias are identified in the research. Clinicians must weigh the pros and cons of each measure in choosing the most appropriate test for their clients.

Of particular importance is an analysis of the strides that have been made by each cultural

measure and the areas that need further development. The measures reviewed above emphasize the importance of specific research on each measure that would highlight issues affecting the validity of test interpretation such as: the need for different cut scores (Heaton, et al., 2003) and effects of culture-specific content and language demands (Kaufman & Lichtenberger, 2006). Well-researched measures such as the EIWA-III (Weschler, 2008) and the Spanish WAIS-III (TEA Ediciones, 2001) take appropriate steps to explore the psychometric properties of the test when applied to a specific population of interest and follow ITC recommendations (Hambleton, 2001). This type of research is of the utmost importance and its implications for diagnostic measures of psychopathology are discussed below. Psychometric properties and important cultural issues specific to the Spanish SIRS are discussed in later sections.

Diagnostic Measures of Psychopathology

As with measures designed to assess intelligence, the argument for different cut scores to use with minority groups extends into the realm of diagnostic measures for psychopathology, particularly for individuals whose primary language is Spanish. When comparing the mean scores of Hispanic Americans and European Americans even on English versions of multiscale inventories, culturally specific response patterns emerge.

Research on the MMPI-2 has consistently found significant "L" scale elevations among Hispanic Americans when compared to European Americans (Callahan, 1998; Campos, 1991). The L scale was developed to detect attempts by patients to present themselves in a favorable light (Hathaway & McKinley, 1989). Elevated patterns suggesting that Hispanic Americans distort their self-reports to appear less impaired are not confined to one measure. Studies looking at the PAI find similar results. Regarding this issue, Hopwood, Flato, Ambwani, Garland, and

Morey (2009) found that Hispanic American participants scored higher than European Americans on all socially desirable response measures used in the study. On this same point, Romain (2000) found that more than 40% of the PAI protocols from Hispanic Americans were considered "invalid" based on the standard cut scores outlined in the PAI manual (Morey, 2007), as compared to 20% of the European American profiles. As a contributing factor, Hispanic Americans had higher Positive Impression Management (PIM) scores when compared to European Americans.

Findings about impression management and socially desirable responding might lead practitioners to surmise that Hispanic Americans are largely reticent to disclose their psychological issues in the formal context of an evaluation and, perhaps, this is why no other diagnostic patterns are sometimes evident on the clinical scales of these particular assessment measures. Hesitation to disclose symptoms might reflect an issue in response style and interview behavior for this population rather than indicate an absence of symptoms (Correa & Rogers, in press). However, other theories of Hispanic American response styles suggest a different explanation. For example, the phenomenon of Extreme Response Style suggests that individuals of certain cultures, particularly Hispanic and Mediterranean cultures, have a tendency to respond at either the extremely low or the extremely high end of the spectrum when given choices on Likert-type scales in the United States (Hui & Triandis, 1989). It is believed that these individuals consider extreme responses to be more sincere than a "conservative" response located in the middle of a Likert-type scale. The distinction is most evident for individuals within these two cultures in contrast to individuals of Asian cultures, who do tend to respond in the middle of the scale (Zax & Takahashi, 1967). Notably, the language of a test can magnify this cultural response style. In a study that administered the same items in two different languages to

bilingual individuals, Gibbons, Zellner, and Rudek (1999) found that participants used more extreme ratings when responding in Spanish than in English. Contrary to research stating that Hispanic Americans tend to respond defensively to multiscale inventories, studies of Extreme Response Styles suggest that extreme responding is possible in *both* directions (i.e., underreporting and overreporting).

A study by Romain (2000) casts further doubt on the assertion that defensiveness is a predominant response style for Hispanic Americans. Despite finding a higher PIM score for Hispanic Americans, Romain (2000) noted that both Hispanic and European Americans showed relatively little withholding or defensiveness as demonstrated by low mean PIM scores of 45.32 and 38.06 respectively. In general, research on the cultural response styles on the PAI is lacking. The normative samples included in the PAI manual create three major limitations in interpreting results for Hispanic American patients. First, ethnic differences for Hispanic Americans are explored in the test manual for the census-matched standardized sample but were not considered for the clinical sample. A second major limitation was the collapsing of all minority groups except African Americans into a single "other" group (see Romain, 2000; Todd, 2005). The clinical standardization samples described in the more recent version of the PAI manual (Morey, 2007) are composed of 78.8% European Americans, 12.6% African Americans, and 8.6% "other" minority groups. Combining all minority groups into a single category does not allow for specific comparisons between groups and it implicitly makes the erroneous assumption that all minority groups are alike, except for African Americans. Thus, this grouping also creates a third major problem by masking minority differences. For instance, high scores for Hispanic Americans on a particular scale might be balanced by low scores from another culture (Correa & Rogers, in press).

Published research conducted with clinical samples has not systematically attempted to identify differences in response patterns of ethnic minority populations. Greene (2000) points out that very little research has examined differences between Hispanic and European Americans on both clinical and validity scales of the MMPI-2. With most of the research having been conducted on undergraduate students with presumably low levels of psychopathology, Greene cautions against making general statements about the cultural response styles of Hispanic American patients on the MMPI-2, concluding that it is premature for this clinical population and that further research is necessary.

A recent study using the Spanish language PAI takes an important first step in evaluating malingering among Spanish-speaking populations. In a within-subjects design Fernandez, Boccaccini, and Noland (2008) used a non-clinical sample of bilingual individuals to assess the performance of PAI validity scales across both language versions. They found that the validity scales, generally, performed similarly in both language versions, with the NIM and PIM scales demonstrating the highest levels of equivalence. Results also indicated possible defensiveness within the sample, as individuals responding honestly exhibited a greater tendency to underreport symptoms on the Spanish version. The authors advise that their results should be interpreted with caution, as their sample of bilingual individuals is different than most samples of monolingual Spanish speakers in levels of acculturation and education.

Given the scarcity of normative data for many assessment measures and minority populations, a primary goal of the current study was to provide as much comprehensive data as possible on the Spanish SIRS and Hispanic Americans. The following section discusses validity and reliability data as well as specific response patterns for Hispanic Americans on the Spanish

SIRS. Comparisons are also made between Hispanic American results in this study and the normative sample of European Americans on the English language version of the SIRS.

The Spanish SIRS and Hispanic Americans

This thesis is the first study to investigate the effectiveness of the Spanish SIRS in distinguishing between feigners and honest responders. Using a Spanish-speaking, predominantly monolingual, sample of Hispanic American outpatients, its overall results indicate that the Spanish SIRS is generally successful and serves as a valuable tool in the assessment of malingering among Spanish speaking individuals. The psychometric properties of the Spanish SIRS are critically examined below.

Reliability

The English language version of the SIRS is considered the gold standard in malingering assessment because of its exceptional reliability, validity, and classification accuracy (Blau, 1998; Lally, 2003). This study found high reliability, validity, and classification accuracy for the Spanish SIRS. Comparable to the English version, whose primary scales exhibited high alpha coefficients (M = .86; range from .77 to .92) the alpha coefficients for the Spanish SIRS in this sample are also generally high (M = .89; range from .76 to .96). Such high alpha coefficients could indicate that items in the individual scales very closely measure the same construct; however, with alpha coefficients as high as .96, there is a possibility that the items on certain primary scales are more redundant when used for assessing Spanish-speaking Hispanic Americans. The strongest alpha coefficients for this sample were found in scales that utilize amplified detection strategies: BL (α = .96) and SU (α = .95). The BL and SU scales involve the

proportions of major symptoms (BL) and everyday problems (SU) reported by an individual. According to Rogers et al. (1992), these two primary scales also exhibited the highest alphas in the original English validation sample (for BL α = .92; for SU α = .92).

RO is the only scale, which produced a moderate alpha (α = .76). This scale also produced the lowest alpha in the original English-language validation sample (α = .77) (Rogers, et al., 1992). Because evaluators must combine their own clinical observations with the participants' self-report on RO, it is interesting that RO exhibited a lower internal consistency than the rest of the Spanish SIRS scales and was also the only subscale that exhibited a smaller effect size for this sample when compared to the original English language validation group. Though it still had a large effect size with the Spanish-speaking sample (d = 1.38), relative to the other primary scales, RO appears to be the least reliable and effective for this population. This is not the case with the English language validation sample, where two other primary scales had relatively lower effect sizes than RO (Rogers, 2008). For the Spanish SIRS, RO had the smallest means and standard deviations of any primary scale for both honest (M =.62, SD = 1.18) and feigning participants (M =3.73, SD = 2.94), so it is possible that the scale may have a floor effect with this population.

Interrater reliability was excellent for each primary scale of the Spanish SIRS, with rs ranging from .98 to 1.0. High interrater reliabilities are not unexpected from a fully structured interview with yes/no responses, such as the Spanish SIRS. Since the publication of the original SIRS manual, research studies routinely report interrater reliabilities for the English version that range from .93 to 1.00 (Rogers, 2008). Notably, for the Spanish SIRS, even the scale that requires individual evaluators to use their own clinical judgment achieved a very high interrater reliability (RO r = .98). Because of the possible floor effect of the RO scale found in this

experimental sample, future research using different populations should also investigate the interrater reliability of RO.

SEM, a basic requirement of scale validation and subsequent interpretation (Anastasi, 1988), is higher for the Spanish SIRS than for the original English language version. Most Spanish SIRS scales have an SEM of about one point (M = 1.06). Overall, these scores indicate a good reliability of individual scores for the Spanish SIRS. However, a recent study using a large database of English language SIRS protocols, found an even smaller average SEM of .29 (Rogers, Payne, Berry, & Granacher, 2009).

Validity

Large effect sizes are crucial for establishing the discriminant validity of the Spanish SIRS. Specifically, large effect sizes are found between the experimental conditions in this study would indicate that feigners and honest responders have measurably different scores on the Spanish SIRS and the measure strongly distinguishes between them. Results from this simulation design indicate that the Spanish SIRS produced very large overall effect sizes when distinguishing feigning participants from honest responders ($M\ d$ = 2.00). Such effect sizes are slightly larger than simulation research in the original validation ($M\ d$ = 1.74; Rogers et al., 1992) for the English language SIRS. Notably, effect sizes for the primary scales of the Spanish SIRS are also larger than the effect sizes noted for other English language measures with similar detection strategies for the assessment of feigning (see Jackson et al 2005; Rogers, 2008; Rogers et al., 2003; Sellbom & Bagby, 2008): the MMPI-2 ($M\ d$ = 1.31), and the PAI ($M\ d$ = 1.45). As subsequently discussed, the differences may be partly due to these measures' reliance on unlikely strategies which—at least in the current study—yielded somewhat lower effect sizes.

While direct comparisons can be made between effect sizes from the Spanish SIRS and those of other English language measures of malingering, unfortunately, no direct comparisons can be made between the Spanish SIRS and other Spanish language measures of feigning. As noted in the introduction, no published studies have investigated malingering in a Spanish-speaking Hispanic American clinical population. In a study using a mixture of clinical and non-clinical Spanish-speaking adolescents in Mexico, Lucio, Duran, Graham, and Ben-Porath (2002) found that four scales (F, F1, and F2 scales, and F-K index) on the Mexican version of the The Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A; Lucio, 1998) adequately discriminated between feigners and honest responders. However, they caution against applying the findings from this study to Hispanic adolescents from the United States. They highlight that cultural differences between adolescents in Mexico and Hispanic Americans in the United States may require different cut-scores. Specifically, Lucio, et al. (2002) have noted that Hispanic American adolescents in the United States tend to be less forthcoming when reporting symptoms than adolescents in Mexico.

The current investigation attempted to avoid unfounded generalizations for the Spanish SIRS, both within and between cultures. For the former, cultural differences were explored by considering participants on the basis of their ARSMA-II level of cultural identification. As could be expected, this effort was only partially successful because a substantial 76.3% of the sample had a Traditional orientation according to the ARSMA-II, indicating little cultural heterogeneity among participants. For the latter, the Hispanic American sample in this study was also contrasted with the original normative sample for the English language version of the SIRS. Overall, there were very large effect sizes for individuals identifying with non-American culture. In most cases, the effect sizes are larger than in the original validation sample of the SIRS, with

RO being the only exception. The SC, IA, and BL scales were each substantially larger in the current sample than in the original validation sample. IA and SC are both considered "unlikely" detection strategies and BL is considered an "amplified" detection strategy. As noted in Table 14, Spanish SIRS scales using amplified detection strategies (i.e., BL, SU, SEL, and SEV) produced somewhat higher effect sizes (*M d* = 2.19 versus *M d* = 1.80) than those utilizing unlikely detection strategies (i.e., RS, SC, IA, and RO) for Spanish-speaking Hispanic Americans. Interestingly, amplified detection strategies also showed relatively higher effect sizes (*M d* = 1.90) in the original validation sample than unlikely detection strategies (*M d* = 1.57), though the differences were not as pronounced. However, differences in average effect size of amplified detection strategies between this sample and the English language validation could be partly due to cultural factors. When comparing Traditional participants to those with other levels of acculturation (see Table 13), those with a traditional orientation exhibited a slightly higher average effect size (*M d* = 2.13 vs *M d* = 2.01) for amplified detection strategies.

The much lower effect sizes for unlikely detection strategies in the current sample might also be attributed to cultural factors. These findings could indicate that Hispanic American individuals have more difficulty identifying symptoms that European American individuals consider to be uncommon or unlikely, making them less prone to endorse these items when attempting to malinger. Alternatively, smaller effect sizes for unlikely detection strategies could reflect some level of defensiveness—even in the feigning condition—or a reticence to endorse symptoms of extreme pathology, even when attempting to feign complete impairment. In either case, unlikely detection strategies appear to be slightly less effective for this population.

Supplementary SIRS Scales

Overall, results provide strong evidence that the Spanish SIRS primary scales and most supplementary scales clearly differentiate between feigned and genuine conditions among Spanish-speaking Hispanic outpatients. The feigning group attained significantly higher scores than the honest group on every supplementary scale of the Spanish SIRS except for INC. Notably, INC demonstrated no significant difference between feigners and honest responders. The INC scale looks at whether the person being evaluated is responding in an inconsistent manner. According to Rogers et al. (1992), it is difficult for malingerers to remember which symptoms they have falsely endorsed and, therefore, often appear inconsistent in their responses. The fact that the INC scale does not appear to differentiate between feigners and honest responders could mean (a) that this population puts increased effort into maintaining consistent responses, even when providing false or exaggerated responses, (b) they did not perceive inconsistency as related to psychopathology, genuine or feigned.

Because researchers have expressed concern regarding defensive response styles exhibited by Hispanic Americans on measures, such as the MMPI-2, MMPI-A, and PAI (Callahan, 1998; Campos, 1991; Lucio, et al., 2002; Romain, 2000), it is important to consider possible defensiveness on the Spanish SIRS. The DS scale measures defensiveness and the denial of everyday problems. Of the Spanish SIRS supplementary scales which showed significant differences between feigners and honest responders, DS had the smallest effect size (d = 1.17). The SIRS manual (Rogers et al., 1992) states that a score lower than 15 points on this scale is indicative of minimization or denial of common problems. For this study, honest responders attained an average scale score higher than 15 points; this would not seemingly indicate defensiveness (M = 18.41). However, 30% of honest participants scored below 15%. As

expected, feigning participants attained a much higher average score (M = 29.10). Only 7.5% of feigning participants scored below 15 points on DS. Romain (2000) also noted that her sample of Hispanic Americans showed relatively little withholding or defensiveness on the PAI when compared to European Americans, suggesting that defensiveness may not be a cultural response style that significantly affects scores on all assessment measures.

Classification Accuracy

Before discussing classification accuracy of the Spanish SIRS, a brief review of SIRS scoring interpretation is helpful. The basic determination of feigning on the SIRS involves (1) one or more primary scales in the definite feigning range, or (2) three or more primary scales in the probable feigning range, or (3) for marginal cases (e.g., one or two scales in the probable feigning range), the SIRS total score > 76 can be employed (Rogers et al., 1992). When these rules were applied to the Spanish SIRS, the overall classification rate was high at .88. Sensitivity (.90) and specificity (.85) were well balanced, and similar estimates were found for positive predictive power (PPP = .86) and negative predictive power (NPP = .89).

A hallmark of the original English language SIRS is its very small false-alarm rate (Rogers et al., 1992). The false-alarm rate in the current study is appreciably higher at 15%. It is unknown whether this difference in false-alarm rate is due to cultural factors or other variables inherent in the current sample because our analyses were limited by the uneven numbers of participants for each level of acculturation. Group size for this sample only allowed us to explore whether the utility estimates would remain stable when applied only to traditionally-oriented Hispanic outpatients. When computed for the traditional group alone, the false-alarm rate remained similar at 14%, and PPP (.88) and NPP (.89) were slightly higher. The English

language version of the SIRS shows no significant differences on the measure due to race (Connell, 1992). However, this sample only investigated differences between European Americans and African Americans. More recent studies examine responses of Hispanic American clinical populations in addition to European and African Americans and find no significant differences in the utility of malingering measures across ethnic groups (Guy & Miller, 2004; Miller, 2005). However, these studies focus on the utility of the M-FAST, rather than the SIRS and stipulate that although the M-FAST was developed using the SIRS as the criterion, it should not be assumed that both measures are affected by cultural variables in a similar manner (Miller, 2005).

Acculturation

In psychological assessment, issues of acculturation must be considered for individuals whose primary identification is toward a different culture (i.e., the traditional orientation, as classified by the ARSMA-II). Researchers and practitioners both recognize that standardized assessment measures administered to individuals who are culturally different from the normative sample can have quite different psychometric characteristics and lead to biased results as well as incorrect classification of individuals from different cultural groups (Marin & Marin, 1991; Dana, 2005). Culturally biased assessment results occur, in large part, because interpretive norms developed mostly on individuals of European American heritage remain valid for only the European American culture if no further testing is conducted on other cultures (Berry 1969, 1988, 1989; Dana, 2005). Culturally-specific response patterns are noted among Hispanic American individuals throughout assessment literature (Campos, 1989; Helms, 1992; Molina & Franco, 1986). Omitting analysis of cultural variables in test development effectively forces

minority individuals into the same interpretative categories as European Americans and creates a substantial possibility for misdiagnosis and misinterpretation (Dana, 1993; Todd, 2004).

In order to avoid inappropriately making generalizations about different cultural groups present in the sample, the current study evaluated possible effects of acculturation on the Spanish SIRS. The sample was divided into different groups, based on their ARSMA-II classifications. This practice is advisable because researchers find that English language measures adapted for Spanish speakers often fail to evaluate level of acculturation (Echemendia & Harris, 2004; Salazar, Perez-Garcia, & Puente, 2007; Renteria et al, 2007). As previously noted, Lucio et al. (2002) point out the detrimental effects of failing to acknowledge cultural differences. In their study of the MMPI-A and Mexican adolescents, they state that cultural differences in response styles likely call for different cut-scores when the same measure is used for American adolescents of Hispanic descent.

The current study used correlations between level of acculturation and performance on Spanish SIRS primary scales to determine if a relationship existed between scale scores and levels of acculturation. All correlations between acculturation and scale scores were non-significant, with the sole exception of the IA scale. IA demonstrated a moderate relationship between scale score and level of acculturation for participants in the honest condition. Within the Honest condition, traditionally oriented participants achieved very low scores on IA (M = .38; SD = .78), while non-traditional participants achieved higher scores on average (M = 2.00, SD = 2.69). This difference could indicate an increased reticence among traditional honest responders to endorse the most absurd symptoms on the Spanish SIRS. In fact, scales employing unlikely detection strategies had lower scores, on average, (M = 1.03) among traditional honest responders than non-traditional honest responders (M = 1.92). While the moderate relationship

between culture and IA is interesting, no other correlations are significant. The general absence of significant correlations appear to suggest that level of acculturation and SIRS scores are largely unrelated. These findings echo results found by Connell (1992) regarding ethnicity and the SIRS, where no significant differences were found between African Americans and European Americans.

Despite an attempt to analyze various levels of acculturation, the current study was only able to adequately analyze results from one cultural group. Due to uneven group size, we were largely limited to studying effect sizes of the Traditionally-oriented group. As summarized in Table 14, we analyzed effect sizes for the Traditional group in both the honest and feigning conditions and then compared their effect sizes with those of the total Spanish-speaking sample and original English language validation group. The overall values were quite similar between the Traditional group and the total sample (M = 2.00 and M = 2.00, respectively). However, these similarities are expected, given the overlap in participants between the two groups (61 of 80 or 76.3% was classified as Traditional), and shed no light on the differences that may be present between different levels of acculturation. There does appear to be differences between the traditional participants and non-traditional participants, however (see Table 13). The traditional group exhibited larger effect sizes on Spanish SIRS primary scales ($M\ d$ = 2.01) than the non-traditional group ($M\ d$ = 1.77). The original English Language validation sample had a similar mean effect size to the non-traditional group with a mean d = 1.74. These results could indicate a culturally specific trend, with increased identification with traditional Hispanic culture, yielding higher effect sizes on the Spanish SIRS. However, potential findings on the effects of acculturation might be obscured due to the lack of cultural variability in the current sample.

Potential Test Bias

The results of this study identify only one potential area of test bias for the Spanish SIRS. Method bias described by Van de Vijver, and Hambleton, (1996) appears to be present in the Spanish SIRS items that require persons to quickly rhyme words or provide the opposites of words read by the evaluator. Efforts were, indeed, made to eliminate this problem during the translation process. Rather than provide a direct translation of these English items, some of the original rhyming and opposite words were changed for the Spanish version. This decision was made because several of the items had no rhyming counterparts in Spanish, making the task impossible. Researchers agree that a direct translation must, at times, be sacrificed so that test items can more accurately convey the intended *message* (Jeanrie & Bertrand, 1999; Marin & Marin, 1991; Solano-Flores, Backhoff, & Contreras-Niño, 2009; Van de Vijver, & Hambleton, 1996). Following this model, words that could be rhymed in Spanish were provided for the Spanish SIRS, effectively removing item bias for this task.

However, method bias inherent in the items involving rhyming and opposite words remained for this experimental sample. The large number of participants made errors on these items and requested to have the task explained to them several times. As a possible explanation, poor participant performance on this task could be due to lower levels of education found in the sample. A substantial 48.8% of the sample had less than a high school education while only 11.4% had greater than a high school education. Furthermore, an overwhelming 88.5% of participants were foreign-born and likely educated outside of the United States, where such rhyming tasks are not a common practice in school systems. Contrary to this explanation, participants with a High School education or lower actually made an average of less errors on these tasks ($M = 5.60$) than those with higher levels of education ($M = 6.32$, $p = .117$). Thus, an

alternative explanation is that low scores on these tasks could be attributed to acculturation, as 76.3% of the sample had a Traditional orientation toward Hispanic culture. The inability of certain cultural groups to perform highly on some cognitive tasks has been widely noted by researchers that point out discrepancies in educational practices between different countries (Artiola, Fortuny, Heaton, & Hermosillo, 1998; Renteria, 2005). Though not statistically significant (p = .60), for this sample, the traditionally oriented group actually made a lower number of errors (M = 5.68) than the non-traditional group (M = 6.33) on the rhyming and opposite tasks. This could be because individuals that more closely identify with traditional Hispanic culture are actually better able to complete these tasks in the Spanish language than individuals who do not identify closely with traditional Hispanic culture. For both cultural groups, errors were more prevalent on the opposite items (M = 2.91) than rhyming items (M = 1.68). Fortunately, these items do not factor into the classification of malingering, so participants' overall Spanish SIRS classification was not affected.

Effects of Psychopathology on Spanish SIRS Classification

A final area of exploration for this study was whether Axis I psychopathology had an effect on Spanish SIRS classification. For the English SIRS, certain forms of severe psychopathology such as PTSD and dissociative disorders have been found to inflate the false-alarm rate (Rogers, Payne, Correa, Gillard, & Ross, 2009). Hence, it is important to explore the role of severe psychopathology for the Spanish SIRS, as well. The information discussed in this section is largely exploratory, as this is the first study to explore malingering in a clinical population using a Spanish language measure. In addition, the measure of psychopathology, the MINI, is intended as a diagnostic screen, and the Spanish MINI remains unvalidated.

Looking at the most prevalent diagnostic categories found in the current sample (Major Depressive Disorder, Generalized Anxiety Disorder, and Psychotic symptoms), weak to moderate correlations were found between diagnostic groups and several primary scales of the Spanish SIRS. For example, GAD showed weak to moderate correlations with several scales using amplified detection strategies: SEL, SEV, and SU. In addition, psychotic symptoms showed moderate correlations with two scales using unlikely detection strategies: RS and SC. This finding is not unexpected; individuals with psychotic symptoms show a relationship to scales with unlikely detection strategies, because many of the items on these scales include bizarre or psychotic content (Rogers et al., 1992). Finally, the MDD category exhibited the largest number of significant correlations (i.e., weak to moderate correlations with SC, IA, BL, SEV, SU, and SEL). The reason for the increased number of significant correlations for MDD is unknown. However, it could be influenced by a several factors (a) as discussed below, there was a large amount of comorbidity evident in this sample, or (b) individuals with MDD in this sample had higher levels of impairment. The aforementioned SIRS study by Rogers et al. (2007) used the Schedule of Affective Disorders and Schizophrenia – Change Version (SADS-C; Spitzer & Endicott, 1978) to determine symptom severity of their participants. The current study used the MINI, which contains no measure of symptom severity, thus no data on participants' level of impairment are available.

Because symptom severity was uncertain, the study looked at whether classification accuracy was affected by symptoms of the diagnostic clusters mentioned above. False alarm rates are notably higher for participants belonging to the aforementioned diagnostic groups (see Table 18). Because of the high rate of comorbidity within the sample, it is possible that these high false-alarm rates were the result of counting the same misclassified individuals multiple

times. In order to remove this confound, individuals with comorbid MDD and GAD symptoms were removed from the sample and false-alarm rates were studied again. False-alarm rate for the MDD only group remained slightly higher than before (see Table 19), while the GAD only group had a lower false-alarm rate than the GAD comorbid group (17.0% vs 25.0%). It should be noted that, removing comorbidity from the diagnostic groups (e.g., psychotic category) left group size unacceptably small and not much stock should be placed on these numbers. Due to widely discrepant group sizes and notable overlap among diagnostic groups, it is not appropriate to use this sample to establish general trends in the relationship between psychopathology and scale scores on the Spanish SIRS. Moreover, it is impossible to discern whether these observed differences reflect on limits in the Spanish MINI rather than the Spanish SIRS.

To date, there are no published research studies on the effects of severe psychopathology on other Spanish language measures with validity scales such as the Spanish PAI or Spanish versions of the MMPI. As observed before, there remains a dearth of malingering research involving Spanish language measures.

Summary

In line with the ITC test guidelines, translations should not be used for clinical evaluation until validated for their intended purpose and target population (Hambleton, 2001). The Spanish SIRS was created using a rigorous back-translation procedure recommended by most researchers (Matias-Carrelo et al., 2003; Marin & Marin, 1991) and this study sought to extablish its psychometric properties for use in the assessment of malingering for Spanish-speaking Hispanic Americans.

Throughout different domains of psychological assessment, there are few Spanish language measures that have been adequately researched and validated for use with Spanish-speaking Hispanic American populations. Currently, no specialized measures of malingering exist for use with these populations. Spanish-language measures with validity scales for whom published validation data are currently available (i.e., MMPI-2 and PAI) have, thus far, neglected to include analyses of these validity scales and associated response styles, such as malingering in adult clinical populations (Correa & Rogers, in press; Fernandez, et al., 2008; Lucio et al., 2002; Romain, 2000). Because the classification of malingering often has important implications for how clinical patients are treated (Rogers & Schuman, 2005), this study sought to provide data on the reliability and validity of a specialized malingering measure for a population that lacked such an assessment tool.

By and large, results from the current study indicate the Spanish SIRS is a useful and valid measure for the classification of feigning in the target population. Positive predictive power (PPP) and Negative predictive power (NPP) were high, and false-alarm rates were low. The alphas and inter-item correlations showed good scale homogeneity for the primary scales of the Spanish SIRS. Furthermore, each primary scale demonstrated very large effect sizes, indicating that the Spanish SIRS is quite effective in differentiating feigners from honest responsers. Lastly, interrater reliabilities are strong and also generally consistent with the original English validation (Rogers et al., 1992). Overall, results indicate the Spanish SIRS is useful for the classification of feigning.

Limitations

Unfortunately, language equivalence could not be tested in this study, as the sample was

largely monolingual. Thus, no direct comparisons can be made about the Spanish and English language versions of the SIRS. An initial test of linguistic equivalence between the English and Spanish versions of the SIRS conducted by Rogers et al. (2009) found very similar scores between both versions for bilingual clients in an outpatient setting in Miami. Their preliminary data indicates that both versions are appropriate for use with bilingual Hispanic American individuals, and clinicians can allow patients to choose their preferred language for testing.

As expected with a monolingual Hispanic American sample, a notable limitation of the study is that there was very little variability in level of acculturation among participants. In fact, it was initially proposed that participants from the control condition would be divided into two groups for data analysis. These two groups were to reflect their level of acculturation based on the scores they attained on the ARSMA-II: Traditional and Assimilated. This analysis would have allowed for comparisons of the most extreme groups, highlighting the most salient differences relating to level of acculturation. However, once data collection was completed, it became apparent that there was not sufficient statistical power to conduct these analyses for each of the acculturation groups, as the vast majority (76.3%) of participants fell into the Traditional group and only 2.5% fell into the Assimilated group. Due to this limitation, the potential effects of acculturation could not be fully explored.

As previously described, one final limitation of the current study is the use of the MINI. Although MINI demonstrated excellent interrater reliability in both English and French (Lecrubier, Sheehan, Weiller, Amorim, Bonora et al., 1997; Sheehan et al., 1998) and its reliability has been maintained in Japanese and Italian translations (Otsubo, Tanaka, & Koda, 2005; Rossi, Alberio, & Porta, 2004). The English language version of the MINI has several drawbacks in a clinical setting. A study by Black, Arndt, Hale, and Rogerson (2004) used the

English language MINI as a screening tool for a prison population. It criticized the adequacy of the measure, emphasizing that (a) the measure does not take symptom severity into account, (b) malingering is not addressed, and (c) staff felt that a number of the modules were confusing and difficult to interpret. An additional drawback to using the MINI is that no published psychometric information on the Spanish version of the MINI exists. Despite these concerns about the MINI, we chose to include it in the current study because concerns about reading ability and level of education in this particular sample prevented the use of written measures. As noted, a substantial 48.8% of the sample had less than a high school education while only 11.4% had greater than a high school education. Thus, we opted to use an interview-based screen, rather than a Spanish language multiscale inventory with published data.

Future Directions

Due to the cultural homogeneity of our sample, it is unknown whether different levels of acculturation have a significant effect on the utility of the Spanish SIRS and on participant scores for the rhyming and opposites tasks. Future research with a more culturally diverse sample of Hispanic Americans can shed light on this area (Salazar et al., 2007). As previously observed, there are no published data on the utility of the Spanish PAI or Spanish versions of the MMPI for the detection of malingering. This major oversight creates another area for future research to explore (Correa & Rogers, in press; Lucio et al., 2002). Since the false-alarm rate in this study was appreciably higher than the English language version of the SIRS, additional research with more varied levels of acculturation is necessary for analysis of more appropriate Spanish SIRS cut scores. The current findings corroborate the conclusion of the International Test Commission

that all psychological measures, including the Spanish SIRS, must be independently validated for each language translation.

Future studies should also be conducted with Hispanic American participants that have higher levels of education. Such studies should make sure to investigate participants' performance on the rhyming and opposites tasks, in order to investigate potential test bias and any emic qualities of this task so that they may be modified.

Lastly, a chi squared test of independence was originally planned to determine the relationships between diagnostic groups (as categorized by the MINI) and the most commonly endorsed items on any Spanish SIRS scales showing moderate correlations. Unfortunately, due to the random assignment of participants to each experimental condition, group size for the diagnostic categories was not large enough to analyze this among honest participants. Future research with larger samples can, hopefully, explore this area. Alternatively, future studies can replace the MINI with an interview-based measure with published validation data such as the Spanish Diagnostic Interview Schedule (Spanish DIS; Burnam, et al., 1983). Upcoming studies can also focus on exploring the effect of psychopathology and the Spanish SIRS in a different manner. Research conducted in an environment where reading level is not a concern could use the Spanish PAI or MMPI-2 to investigate these variables. Another benefit to using these multiscale inventories is that a comparison of effect sizes can be made for the detection strategies used by their validity scales and corresponding detection strategies on Spanish SIRS scales.

APPENDIX A

INFORMED CONSENT FORM

University of North Texas

Institutional Review Board

Informed Consent Form

Before agreeing to participate in this research study, it is important that you read and understand the following explanation of the purpose and benefits of the study and how it will be conducted.

Title of Study: Validation of the Spanish SIRS: Beyond Linguistic Equivalence in the Assessment of Malingering Among Spanish Speaking Clinical Populations

Principal Investigator: Amor Correa, a graduate student in the University of North Texas (UNT) Department of Psychology.

Purpose of the Study:

You are being asked to participate in a research study which involves talking about psychological symptoms, and how role-played conditions affect how symptoms are reported.

Study Procedures:

Through interviews and questionnaires, you will be asked to answer questions about the experience of psychological symptoms. You will be asked to complete a total of 4 measures (3 brief interviews and one questionnaire). Some participants will complete the measures under standard instructions; other participants will be asked to role-play a different set of psychological problems. Without rushing, this will take slightly more than one hour of your time. You can also have breaks.

Foreseeable Risks:

The foreseeable risks are negligible. It is possible that you may find a few questions to be minimally distressful. Please let the researcher know if this happens.

You will not be asked whether you have engaged in child abuse or elder abuse. If you volunteer that you have committed or plan to commit child abuse or elder abuse, we are required by law to inform the authorities. When you are asked to "role-play" a different disorder and problems, we believe this information is invalid. Therefore, we will not report "made-up" problems to the authorities.

Benefits to the Subjects or Others:

You may learn things about yourself from this research. The research may help us to understand how psychological distress can affect patients' responses on these questionnaires. This

information is important for treatment because effective treatments rely on accurate responses to questionnaires.

Compensation for Participants:

All participants who attempt to follow their instructions will receive $10 as compensation for their participation upon completion of all parts of the study.

Procedures for Maintaining Confidentiality of Research Records:

Your information will be kept confidential and the research data will be recorded without names or personal identifiers. Your clinic records will not be read as part of this study; no information from them will be recorded. Your signed consent forms and coded survey results will be kept in separate locations. You agree that researchers can contact the clinical staff if you pose a significant risk of suicide, self-harm, or physical aggression towards others. The confidentiality of your individual information will be maintained in any publications or presentations regarding this study.

Questions about the Study:

If you have any questions about the study, you may contact Amor Correa at telephone number _____ or the faculty advisor, Dr. Richard Rogers, UNT Department of Psychology at telephone number _____.

Review for the Protection of Participants: This research study has been reviewed and approved by the UNT Institutional Review Board (IRB). The UNT IRB can be contacted at _____ with any questions regarding the rights of research subjects.

Research Participants' Rights: Your signature below indicates that you have read or have had read to you all of the above and that you confirm all of the following:

- Amor Correa has explained the study to you and answered all of your questions. You have been told the possible benefits and the potential risks and/or discomforts of the study.
- You understand that you do not have to take part in this study, and your refusal to participate or your decision to withdraw will involve no penalty or loss of rights or benefits. The study personnel may choose to stop your participation at any time.
- You understand why the study is being conducted and how it will be performed.
- You understand your rights as a research participant and you voluntarily consent to participate in this study.
- You have been told you will receive a copy of this form.

_____
Printed Name of Participant

_____ _____
Signature of Participant Date

For the Principal Investigator or Designee: I certify that I have reviewed the contents of this form with the participant signing above. I have explained the possible benefits and the potential risks and/or discomforts of the study. It is my opinion that the participant understood the explanation.

_____ _____
Signature of Principal Investigator or Designee Date

APPENDIX B

DEMOGRAPHICS QUESTIONNAIRE

Identification number: _____

Age: _____

Gender: _____

Ethnicity/Family's Country of Origin: _____

SES: Lower Middle Upper

Education Completed:	Elementary	Middle	High School
	GED	Vocational Training	Some College
	Bachelors	Masters	Doctorate
Employment History: (Last Year)	Unemployed	Part-time	Full-time
	Sick Leave	Disabled	

APPENDIX C

FEIGNING INSTRUCTIONS

Role-Playing Instructions:

Your Goal:

For this study, please role-play a person who is pretending to be completely disabled. Assume you are a hard worker and have been doing well at your job for 5 years. This is a good job that gives you very good insurance and disability benefits. Imagine that the new boss doesn't like you and threatens to fire you. If you pretend to be totally disabled, your insurance will pay you $1,000 per month and you won't get fired. You have to convince the insurance company that you have a severe mental disorder that prevents you from working at your job.

Although this is only for a research experiment, please try to be as convincing as possible. It may sound easy, but the hard part will be convincing the interviewer that you are completely disabled by the symptoms that you are pretending to have.

Your Reward:

Can you fool the examiner? These tests are made to catch people who are trying to fake a mental disorder. Are you clever and convincing enough to avoid getting caught? You will receive $10.00 for being successful.

Before beginning the study, please take a moment to think about how you will answer the questions to appear disabled. You will be asked about this later.

APPENDIX D

HONEST INSTRUCTIONS

Accurate Presentation of Symptoms:

Your goal:

Please be open and honest in describing your symptoms and circumstances. Your job is to provide an accurate presentation of your current symptoms and psychological concerns.

Importance:

Please take this study seriously. There are not many psychological tests available for people who speak Spanish. Your participation will help us make sure this Spanish language test is useful and accurate when it is used.

APPENDIX E

DEBRIEFING FORM

Debriefing

Research number: _____
Experimental Condition: ___ malingering, ___control

1. What were your instructions? **[record verbatim]** ___correct, ___incorrect

2. What were your incentives?
   Malingering: (smart enough) ___correct, ___incorrect
   Control: (describe accurately) ___correct, ___incorrect
   Both: ($10.00) ___correct, ___incorrect

3. Did you follow the instructions? ___yes, ___no

4. (If yes) How would you describe your effort at following the instructions?
   ___poor, ___average, ___good

5. On a scale from 0 to 100%, evaluate the effort you put into following the instructions.
   _____%

6. Were you aware that there were questions designed to see if you were faking?
   ___yes, ___no

7. Can you give me some ideas on how these questions were supposed to work? **[record verbatim]**

8. [Malingering condition only] How did you try to answer the questions in order to appear completely disabled? **[record verbatim]**

REFERENCES

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, *16*(1), 55-73.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

American Psychological Association. (1993). Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. *American Psychologist, 48,* 45-48.

Anastasi, A (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Artiola, I. Fortuny, L., Heaton, R. K., & Hermosillo, D. (1998). Neuropsychological comparisons of Spanish-speaking participants from the U.S.–Mexico border region versus Spain. *Journal of the International Neuropsychological Society, 4,* 363-379.

Berkanovic, E. (1980). The effects of inadequate translation on Hispanics' responses to health surveys. *American Journal of Public Health, 70,* 1273-1276.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4,* 119-128.

Berry, J. W. (1988). Imposed etics-emics-derived etics: The operationalization of a compelling idea. *International Journal of Psychology, 24,* 721-735.

Berry, D., Baer, R. A. , Rinaldo, J. C., & Wetter, M. W. (2002). Assessment of malingering In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (2nd ed). New York: Oxford University Press.

Berry, J., Kin, U., Power, S., Young, M., & Bujaki, M. (1989). Acculturation attitudes in plural societies. *Applied Psychology: An International Review, 38,* 185-206.

Bersoff, D. N. (Ed.). (2004). *Ethical conflicts in psychology*. Washington, DC: American Psychological Association.

Black, D., Arndt, S., Hale, N., & Rogerson, R. (2004). Use of the mini international neuropsychiatric interview (MINI) as a screening tool in prisons: Results of a preliminary study. *Journal of the American Academy of Psychiatry and the Law*, *32*(2), 158-162. Retrieved from PsycINFO database.

Blau, T. H., (1998). *The psychologist as expert witness* (2nd ed.). New York: John Wiley & Sons, Inc.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry and Law, 21,* 35-76.

Bourg, S., Connor, E. J., & Landis, E. E. (1995). The impact of expertise and sufficient information on psychologists' ability to detect malingering. *Behavioral Sciences & the Law, 13,* 505-515.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Newbury Park, CA: Sage.

Burnam, A., Karno, M., Hough, R. L., Escobar, J. I., & Forsyth, A. B. (1983). The Spanish Diagnostic Interview Schedule: Reliability and comparison with clinical diagnoses. *Archives of General Psychiatry, 40,* 1189-1196.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Campos, L. (1989). Adverse impact, unfairness, and bias in the psychological screening of Hispanic peace officers. *Hispanic Journal of Behavioral Sciences*, *11*(2), 122-135.

Canino, G. J., Bird, H. R., Shrout, P. E., Rubio-Stipec, M., Bravo, M., Martinez, R., et al. (1987). The Spanish Diagnostic Interview Schedule: Reliability and concordance with clinical diagnoses in Puerto Rico. *Archives of General Psychiatry, 44*, 720-726.

Callahan, W. J. (1998). MMPI-2, symptom reports, and acculturation of White- and Mexican-Americans in psychiatric, college, and community settings. *Dissertation Abstracts International, 58(8-B)*, 4439.

Cha, E., Kim, K., & Erlen, J. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing*, *58*(4), 386-395.

Connell, D. (1992, December). The SIRS and the M test: The differential validity and utility of two instruments designed to detect malingered psychosis in a correctional sample. *Dissertation Abstracts International*, 53, Retrieved from PsycINFO database.

Correa, A., & Rogers, R. (in press). Cross-cultural applications of the PAI. In M. Blais, M. Baity, & C. Hopwood (Eds.), *Clinical applications of the Personality Assessment Inventory.* Routledge: New York, NY.

Cuellar, I., Arnold, B., & Maldonado, R. (1995). Acculturation Rating Scale for Mexican Americans-II: A revision of the original ARSMA Scale. *Hispanic Journal of Behavioral Science, 17*, 275-304.

Cunningham, M., & Reidy, T. J. (1999). Don't confuse me with the facts: Common errors in violence risk assessment at capital sentencing. *Criminal Justice and Behavior, 26,* 20-43.

Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.

Dana, R. H. (1995). Culturally competent MMPI assessment of Hispanic populations. *Hispanic Journal of Behavioral Sciences, 17,* 305-319.

Dana, R. H. (2000). *Handbook of cross-cultural and multicultural personality assessment.* Mahwah, NJ: Lawrence Erlbaum.

Dana, R. H. (2005). *Multicultural assessment principles, applications, and examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

DeClue, G. (2002). Practitioner's corner feigning ≠ malingering: A case study. *Behavioral Science and the Law, 20*, 717-726.

DuAlba, L., & Scott, R. (1993). Somatization and malingering for workers' compensation applicants: A cross-cultural MMPI study. *Journal of Clinical Psychology*, *49*(6), 913-917.

Echemendia, R. J., & Harris, J. G. (2004). Neuropsychological test use with Hispanic=Latino populations in the United States: Part II of a national survey. *Applied Neuropsychology, 11(1),* 4-12.

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

Eignor, D. (2001). Standards for the development and use of tests: The standards for educational and psychological testing. *European Journal of Psychological Assessment*, *17*(3), 157-163.

Fantoni-Salvador, P., & Rogers, R. (1997). Spanish version of the MMPI-2 and PAI: an investigation of concurrent validity with Hispanic patients. *Assessment, 4,* 29-93.

Fernandez, K., Boccaccini, M., & Noland, R. (2008). Detecting over- and underreporting of psychopathology with the Spanish-language Personality Assessment Inventory: Findings from a simulation study with bilingual speakers. *Psychological Assessment*, *20*(2), 189-194.

Gamst, G., Dana, R., Der-Karabetian, A., Aragón, M., Arellano, L., & Kramer, T. (2002). Effects of Latino acculturation and ethnic identity on mental health outcomes. *Hispanic Journal of Behavioral Sciences*, *24*(4), 479-504.

García, L., Ruiz, M., & Abad, F. (2003). Factor structure of the Spanish WAIS-III. *Psicothema*, *15*(1), 155-160.

Geisinger, K. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312.

Geller, J. L., Erlen, J., Kaye, N. S., & Fisher, W. H. (1990). Feigned insanity in nineteenth-century America: Tactics, trials, and truth. *Behavioral Sciences and the Law, 8*, 3-26.

Gibbons, J., Zellner, J., & Rudek, D. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research*, *33*(4), 369-381.

Gordon, M. M. (1964). *Assimilation in American life*. New York: Oxford University Press.

Gorman, W. (1982). Defining malingering. *Journal of Forensic Sciences, 27,* 401-407.

Gough, H. (1947). Simulated patterns on the Minnesota Multiphasic Personality Inventory. *Journal of Abnormal and Social Psychology*, *42*(2), 215-225.

Graham, J. R. (1990). *MMPI-2: Assessing personality and psychopathology* (2nd ed.). New York: Oxford University, Inc.

Green, D., Rosenfeld, B., Dole, T., Pivovarova, E., & Zapf, P. (2008). Validation of an abbreviated version of the Structured Interview of Reported Symptoms in outpatient psychiatric and community settings. *Law and Human Behavior*, *32*(2), 177-186.

Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Grow, R., McVaugh, W., & Eno, T. (1980). Faking and the MMPI. *Journal of Clinical Psychology*, *36*(4), 910-917.

Guy, L., Kwartner, P., & Miller, H. (2006). Investigating the M-FAST: Psychometric properties and utility to detect diagnostic specific malingering. *Behavioral Sciences & the Law*, *24*(5), 687-702.

Guy, L., & Miller, H. (2004). Screening for malingered psychopathology in a correctional setting: Utility of the Miller-Forensic Assessment of Symptoms Test (M-FAST). *Criminal Justice and Behavior*, *31*(6), 695-716.

Guy, L., Poythress, N., Douglas, K., Skeem, J., & Edens, J. (2008). Correspondence between self-report and interview-based assessments of antisocial personality disorder. *Psychological Assessment*, *20*(1), 47-54.

Hagglund, L. (2009). Challenges in the treatment of factitious disorder: A case study. *Archives of Psychiatric Nursing*, *23*(1), 58-64.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164-172.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, *11*, 147-157.

Hambleton, R., Yu, J., & Slater, S. (1999). Field test of the ITC Guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*, *15*(3), 270-276.

Hare, R. D. (2003). *Manual for the Hare Psychopathy Checklist—Revised* (2nd ed.). Toronto: Multi-Health Systems.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10,* 249-254.

Hawes, S., & Boccaccini, M. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment*, *21*(1), 112-124.

Heaton, R., Taylor, M., & Manly, J. (2003). Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. *Clinical interpretation of the WAIS-III and WMS-III* (pp. 181-210). San Diego, CA US: Academic Press.

Heilbrun, K., Bennett, W., White, A., & Kelly, J. (1990). An MMPI-based empirical model of malingering and deception. *Behavioral Sciences & the Law*, *8*(1), 45-53.

Helms, J. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing?. *American Psychologist*, *47*(9), 1083-1101.

Hopwood, C. J., Flato, C., Ambwani, S., Garland, B. H., & Morey, L. C. (2009). A comparison of Latino and Anglo positive responding. *Journal of Clinical Psychology, 65*(7), 769-780.

Hui, C., & Triandis, H. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296-309.

Jackson, R. L., Rogers, R., & Sewell, K. W. (2005). Forensic applications of the Miller Forensic Assessment of Symptoms Test (M-FAST): Screening for feigned disorders in competency to stand trial evaluations. *Law and Human Behavior, 29,* 199-210.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. *European Journal of Psychological Assessment*, *15*(3), 277-283.

Jones P. S., Lee J. W., Pillips L. R., Zhang X. E. & Jaceldo K. B. (2001). An adaptation of Brislin's translation model for cross-cultural research. *Nursing Research, 50*, 300-304.

Karno, M., Burnam, A., Escobar, J. L., Hough, R. L., & Eaton, W. W. (1983). Development of the Spanish-language version of the National Institute of Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry, 40*, 1183-1188.

Kaufman, A. S., & Kaufman, N. L., (1990). Kaufman Brief Intelligence Test. Circle Pines, MN: AGS Publishing.

Kaufman, A. S., & Kaufman, N. L., (2004a). Kaufman Assessment Battery for Children (2nd ed.). Circle Pines, MN: AGS Publishing.

Kaufman, A. S., & Kaufman, N. L., (2004b). Kaufman Brief Intelligence Test (2nd ed.). Circle Pines, MN: AGS Publishing.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Kusyszyn, I., & Jackson, D. N. (1968). A multimethod factor analytic appraisal of endorsement and judgment methods in personality assessment. *Educational and Psychological Measurement, 28,* 1047-1061.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice, 34,* 491-498.

Lecrubier, Y., Sheehan, D., Weiller, E., Amorim, P., Bonora, I., Sheehan, K., et al. (1997). The MINI International Neuropsychiatric Interview (M.I.N.I.) A short diagnostic structured interview: Reliability and validity according to the CIDI. *European Psychiatry, 12,* 224-231.

Lucio, E. (1998). Spanish version of the Minnesota Multiphasic Personality Inventory: MMPI-A for Mexico. Mexico City, Mexico: El Manual Moderno.

Lucio, E., Durán, C., Graham, J., & Ben-Porath, Y. (2002). Identifying faking bad on the Minnesota Multiphasic Personality Inventory-Adolescent with Mexican adolescents. *Assessment*, *9*(1), 62-69.

Lucio, E., Reyes-Lagunes, I., & Scott, R. L. (1994). MMPI-2 for Mexico: Translation and adaptation. *Journal of Personality Assessment, 63,* 1, 105-116.

Malcarne, V. L., Chavira, D. A., Fernandez, S., & Liu, P. (2006). The Scale of Ethnic Experience: Development and psychometric properties. *Journal of Personality Assessment 86,* (2), 150-161.

Marin, G., Perez-Stable, E. J., & Marin, B. V. (1989). Cigarette smoking among San Francisco Hispanics: The role of acculturation and gender. *American Journal of Public Health, 79,* 196-198.

Marin, G., & VanOss Marin, B., (1991). *Research with Hispanic populations*. Newbury Park, CA: Sage Publications.

Martinez, G., Marin, B. V., & Schoua-Glusberg, A. (2006). Translating from English to Spanish: The 2002 national survey of family growth. *Hispanic Journal of Behavioral Sciences, 28,* 531-545.

Matias-Carrelo, L., Chavez, L., Negron, G., Canino, G., Aguilar-Gaxiola, S., & Hoppe, S. (2003). The Spanish translation and cultural adaptation of five mental health outcome measures. *Culture, Medicine, and Psychiatry, 27,* 291–313.

Meagher, J. F. (1919). Malingering in relation to war neuropsychiatric conditions, especially hysteria. *Medical Record, 96*, 963-972.

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30,* 525-564.

Meyer, R. G., & Deitsch, S. E. (1996). *The clinician's handbook: Integrated diagnostics, assessment, and intervention in adult and adolescent psychopathology* (4th ed.). Allyn & Bacon, MA: Needham Heights.

Miller, H. A. (2001). *M-FAST: Miller Forensic Assessment of Symptoms Test and professional manual.* Odessa, FL: Psychological Assessment Resources.

Miller, H. (2005). The Miller-Forensic Assessment of Symptoms Test (M-Fast): Test generalizability and utility across race, literacy, and clinical opinion. *Criminal Justice and Behavior*, *32*(6), 591-611.

Molina, R., & Franco, J. (1986). Effects of administrator and participant sex and ethnicity on self-disclosure. *Journal of Counseling & Development*, *65*(3), 160-162.

Moreland, K. L. (1996). Persistent issues in multicultural assessment of social and emotional functioning. In L. A. Suzuki, P. J. Mueller, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment* (pp. 51-76). San Francisco: Jossey Bass.

Morey, L. M. (1991). *The Personality Assessment Inventory professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Morey, L. C. (2007). The Personality Assessment Inventory (2nd ed). Lutz, FL: Psychological Assessment Resources, Inc.

Okazaki, S., & Sue, S. (1995). Methodological issues in assessment research with ethnic minorities. *Psychological Assessment. 7.* 367-375.

Olmedo, E. (1981). Testing linguistic minorities. *American Psychologist*, *36*(10), 1078-1085.

Otsubo, T., Tanaka, K., & Koda, R. (2005). Reliability and validity of Japanese version of the Mini-International Neuropsychiatric Interview. *Psychiatry and Clinical Neurosciences, 59*(5), 517-526.

Overholser, J. (1990). Differential diagnosis of malingering and factitious disorder with physical symptoms. *Behavioral Sciences & the Law*, *8*(1), 55-65.

Paulhus, D. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598-609.

Paulhus, D., Bruce, M., & Trapnell, P. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, *21*(2), 100-108.

Peña, E. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255-1264.

Pons, J.I., Flores-Pabon, L., Matias-Carrelo, L., Rodriguez, M., Rosario-Hernandez, E., Rodriguez, J. M., Herrans, L. L., Yang, J. (2008). Confiabilidad de la Escala de Inteligencia Weschler para Adultos Version III, Puerto Rico (EIWA-III). [Reliability of the Weschler Adult Intelligence Scale, Version III, Puerto Rico [EIWA-III]. *Revista Puertorriqueña de Psicologia [Puerto Rican Psychology Magazine],* 19, 112-132

Pope, C. (1919). Malingering. *New York Medical Journal, 109*, 977-997.

Reid, W. H. (2000). Malingering. *Journal of Psychiatric Practice*, *6*, 226-228.

Renteria, L. (2005). Validation of the Spanish Language Wechsler Adult Intelligence Scale (3rd ed.) in a sample of American, urban, Spanish speaking Hispanics. *Dissertation Abstracts International*, 66, Retrieved from PsycINFO database.

Renteria, L., Li, S., & Pliskin, N. (2008). Reliability and validity of the Spanish Language Wechsler Adult Intelligence Scale (3rd edition) in a sample of American, urban, Spanish-speaking Hispanics. *Clinical Neuropsychologist*, *22*(3), 455-470.

Resnick, P. (1984). The detection of malingered mental illness. *Behavioral Sciences & the Law*, *2*(1), 21-38.

Retzlaff, P., Sheehan, E., & Fiel, A. (1991). MCMI-II report style and bias: Profile and validity scales analyses. *Journal of Personality Assessment*, *56*(3), 466-477.

Rogers, R. (1984). Towards an empirical model of malingering and deception. *Behavioral Sciences and the Law, 2,* 93-112.

Rogers, R. (1990). Models of feigned mental illness. *Professional Psychology: Research and Practice, 21,* (3), 182-188.

Rogers, R. (Ed). (1997). *Clinical assessment of malingering and deception* (2nd ed.). New York: The Guilford Press.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford Press.

Rogers, R. (2008). *Clinical assessment of malingering and deception* (3rd ed.). New York, NY US: Guilford Press.

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Rogers, R., & Bender, S. D. (2003). Evaluation of malingering and deception. In A. M. Goldstein (Ed.), *Comprehensive handbook of psychology: Forensic psychology* (vol. 11, pp. 109-129). New York: Wiley.

Rogers, R., & Cavanaugh, J. L. (1983). "Nothing but the truth" …a re-examination of malingering. *Journal of Psychiatry and Law, 11,* 443-460.

Rogers, R., & Cruise, K. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior*, *22*(3), 273-285.

Rogers, R., Flores, J., Ustad, K., & Sewell, K. W. (1995). Initial validation of the personality assessment inventory—Spanish version with clients from Mexican American communities: A brief report. *Journal of Personality Assessment, 64,* 340-348.

Rogers, R., Gillis, J. R., Bagby, R. M., & Monteiro, E. (1991). Detection of malingering on the SIRS: A study of coached and uncoached simulators. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 673-677.

Rogers, R., Gillis, J. R., Dickens, S. E., & Bagby, R. M. (1991). Standardized assessment of malingering: Validation of the Structured Interview of Reported Symptoms. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 89-96.

Rogers, R., Jackson, R. L., Sewell, K. W., & Salekin, K. L. (2005). Detection strategies for malingering: A confirmatory factor analysis of the SIRS. *Criminal Justice and Behavior, 32,* 511-525.

Rogers, R., Payne, J., Berry, D., & Granacher, R. (2009). Use of the SIRS in compensation cases: An examination of its validity and generalizability. *Law and Human Behavior*, *33*(3), 213-224.

Rogers, R., Payne, J., Correa, A., Gillard, N., & Ross, C. (2009). A study of the SIRS with severely traumatized patients. *Journal of Personality Assessment*, *91*(5), 429-438.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2009). *SIRS professional manual* (2nd ed.). Manuscript in preparation.

Rogers, R., Sewell, K., Martin, M., & Vitacco, M. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment*, *10*(2), 160-177.

Rogers, R., & Schuman, D. W. (2005). *Fundamentals of forensic practice: Mental health and criminal law*. New York: Springer.

Rogers, R., & Vitacco, M. J. (2002). Forensic assessment of malingering and related response styles. In B. Van Dorsten (Ed.), *Forensic psychology: From classroom to courtroom* (pp. 83-104). New York: Kluwer Academic.

Romain, P. M. (2000). Use of the Personality Assessment Inventory with an ethnically diverse sample of psychiatric outpatients. *Dissertation Abstracts International, 61(11-B),* 6147.

Rossi, A., Alberio, R., & Porta, A. (2004). The reliability of the Mini-International Neuropsychiatric Interview-Italian Version. *Journal of Clinical Psychopharmacology, 24*(5), 561-563.

Ryder, A., Alden, L., & Paulhus, D. (2000). Is acculturation unidimensional or bidimensional? A head-to-head comparison in the prediction of personality, self-identity, and adjustment. *Journal of Personality and Social Psychology*, *79*(1), 49-65.

Sackeim, H. A., & Gur, R. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology, 47*, 213-215.

Salazar, G. D., Perez-Garcia, M., & Puente, A. E. (2007). Clinical neuropsychology of Spanish speakers: The challenge and pitfalls of a neuropsychology of a heterogenous population. In B. P. Uzzell, M. Ponton, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 283–302). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Sattler, J. (2001). *Assessment of children cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Sellbom, M., & Bagby, R. (2008). Validity of the MMPI-2-RF (restructured form) L-r and K-r scales in detecting underreporting in clinical and nonclinical samples. *Psychological Assessment*, *20*(4), 370-376.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janvas, J., Weiller, E., et al. (1998). The Mini International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59* (Suppl, 20), 22-33.

Sireci, S., Han, K., & Wells, C. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, *13*(2), 108-131.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations: A methodological analysis of translation quality. *Journal of Cross-Cultural Psychology, 37,* 557 – 567.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. (2009). Theory of test translation error. *International Journal of Testing*, *9*(2), 78-91.

Spitzer, R. L., & Endicott, J. (1978). *Schedule for affective disorders and schizophrenia* (3rd ed.) New York: Biometrics Resarch.

TEA Ediciones. (2001). *WAIS-III Escala de Inteligencia de Wechsler para Adultos–III: Manual tecnico* [WAIS-III Weschler Adult Intelligence Scale-III: Technical manual]. Madrid, Spain: TEA Ediciones.

Todd, W. (2005). Race/ethnicity and the Personality Assessment Inventory (PAI): The impact of culture on diagnostic testing in a college counseling center. *Dissertation Abstracts International, 65(10-B),* 5425.

Turner, M. (1999). Malingering, hysteria, and the factitious disorders. *Cognitive Neuropsychiatry*, *4*(3), 193-201.

US Census Bureau. (2000). Language spoken at home for the citizen population 18 years and over who speak English less than "very well," for the United States, states, and counties: 2000. Census 2000. Retrieved October 13, 2009 from the World Wide Web: http://www.census.gov/population/www/socdemo/lang_use.html.

US Census Bureau. (2004). Hispanic population in the United States: March 2004. *Current Population Survey*. Retrieved October 13, 2009 from the World Wide Web: http://www.census.gov/population/socdemo/hispanic/ASEC2004/2004CPS_tab7.2.txt.

Valencia, R., & Rankin, R. (1985). Evidence of content bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology*, *77*(2), 197-207.

Van de Vijver, F., & Hambleton, R. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89-99.

Vander Kolk, C. (1991). Client credibility and coping styles. *Rehabilitation Psychology*, *36*(1), 51-56.

Vilar-López, R., Santiago-Ramajo, S., Gómez-Río, M., Verdejo-García, A., Llamas, J., & Pérez-García, M. (2007). Detection of malingering in a Spanish population using three specific malingering tests. *Archives of Clinical Neuropsychology*, *22*(3), 379-388.

Vitacco, M., Jackson, R., Rogers, R., Neumann, C., Miller, H., & Gabel, J. (2008). Detection strategies for malingering with the Miller Forensic Assessment of Symptoms test: A confirmatory factor analysis of its underlying dimensions. *Assessment*, *15*(1), 97-103.

Wagner, J., & Gartner, C. G. (1997). Highlights of the 196 institute on psychiatric services. *Psychiatric Services, 48,* 51-55.

Walters, G., Rogers, R., Berry, D., Miller, H., Duncan, S., McCusker, P., et al. (2008). Malingering as a categorical or dimensional construct: The latent structure of feigned psychopathology as measured by the SIRS and MMPI-2. *Psychological Assessment*, *20*(3), 238-247.

Weschler, D. (2008). *Escala de Inteligencia de Weschler para Adultos-Tercera Edicion (EIWA-III)* [Weschler Adult Intelligence Scale-Third Edition (WAIS-III)]. San Antonio, TX: NCS Pearson, Inc.

Wessely, S. (2003). Malingering: Historical perspectives. In P. Halligan, C. Bass, & D. Oakley (Eds.), *Malingering and illness deception* (pp. 31-41). Oxford: Oxford University Press.

Whyte, S., Fox, S., & Coxell, A. (2006). Reporting of personality disorder symptoms in a forensic inpatient sample: Effects of mode of assessment and response style. *Journal of Forensic Psychiatry & Psychology*, *17*(3), 431-441.

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: Comparisons of Japanese and American college students. *Journal of Social Psychology*, *71*(1), 3-10.