Published in final edited form as:
Psychol Assess. 2018 October ; 30(10): 1267–1276. doi:10.1037/pas0000569.

Successful Validation of the CAT-MH Scales in a Sample of Latin American Migrants in the U.S. and Spain

Robert D. Gibbons, PhD,
University of Chicago

Margarita Alegria, PhD,
Disparities Research Unit, Department of Medicine, Massachusetts General Hospital;

Departments of Medicine and Psychiatry, Harvard Medical School

Li Cai, PhD
University of California — Los Angeles

Lizbeth Herrera, MS, **Sheri Lapatin Markle, MIA**
Disparities Research Unit, Department of Medicine, Massachusetts General Hospital

Francisco Collazo, MD,
Department of Psychiatry, Hospital Universitari Vall d'Hebron, CIBERSAM

Department of Psychiatry and Forensic Medicine, Autonomous University of Barcelona

Enrique Baca Garcia, MD
Department of Psychiatry, IIS, Jimenez Diaz Foundation;

Psychiatry Department, Autonoma University of Madrid;

Department of Psychiatry, University Hospital Rey Juan Carlos;

Department of Psychiatry, General Hospital of Villalba;

Department of Psychiatry, University Hospital Infanta Elena;

CIBERSAM (Centro de Investigación en Salud Mental), Carlos III Institute of Health;

Universidad Católica del Maule, Chile

Abstract

We examined cultural differences in the item characteristic functions of self-reported of symptoms of depression, anxiety, and mania/hypomania in a Latino population taking computerized adaptive tests (CAT-MH) in Spanish versus a non-Latino sample taking the tests in English. We studied differential item functioning (DIF) of the most common adaptively administered symptom items out of a bank of 1008 items between Latino (n=1276) and non-Latino (n=798) subjects. For

Correspondence Address: Robert Gibbons, PhD, University of Chicago, Department of Medicine | Department of Public Health Sciences, 5841 S. Maryland Ave., Room W260, MC2000, Chicago, IL, 60637.

Robert D. Gibbons is a founder of Adaptive Testing Technologies which is a company that distributes the CAT-MH suite of computerized adaptive tests. The terms of this arrangement have been reviewed and approved by the University of Chicago in accordance with its conflict of interest policies.

depression, we identified 4 items with DIF that were good discriminators for non-Latinos, but poor discriminators for Latinos. These items were related to cheerfulness, life satisfaction, concentration and fatigue. The correlation between the original calibration and a Latino-only new calibration after eliminating these items was r=0.990. For anxiety, no items with DIF were identified. The correlation between the original and new calibrations was r=0.993. For mania/hypomania, we identified 4 items with differential item functioning that were good discriminators for non-Latinos, but poor discriminators for Latinos. These items were related to risk-taking, self-assurance, and sexual activity. The correlation between the original and new calibration was r=0.962. Once the identified items were removed, the correlation between the original calibration and a Latino-only calibration was r=0.96 or greater. These findings reveal that the CAT-MH can be reliably used to measure depression, anxiety and mania in Latinos taking these tests in Spanish.

Keywords

Depression; anxiety; mania; item response theory; computerized adaptive testing; differential item functioning; Spanish translation

Public clinics screen patients from different cultural backgrounds for mental health conditions, with limited time for a comprehensive assessment and despite concerns that the information collected during a clinical interview might be less discriminating for minority than for majority patients (Vacc & Juhnke, 1997), it is routine practice in both primary care clinics and specialty mental health clinics to use assessments that were developed in the majority population and use them immediately in the minority population following translation. However, patients' perceptions of mental health are socially constructed (Mechanic, 2002, Mojtabai, 2008) and may therefore be interpreted and experienced differently in different populations. There is the belief that inaccurate diagnoses more likely occur when the culture of the patient enters the picture (Malgady & Zayas, 2001; *Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General*, 2001) since culture affects the expression of psychopathology (Favazza & Oman 1984; Kleinman, 1988; Westermeyer & Janca, 1997), as well as diagnosis (Mezzich et al., 1999). As a consequence, we need to understand whether there are differences in the degrees to which symptoms differentiate high and low levels of mental health constructs between Latinos and non-Latinos.

As noted, a common practice in mental health assessment is to take a diagnostic instrument or dimensional severity scale developed in one language and culture, translate it to a different language and then use it in a different culture. Differences in the parameters of the score distributions between the two cultures (e.g. mean and variance) are then interpreted as if they represent differences in the underlying disease or construct of interest. Of course, this assumes that the properties of the administered items are invariant between the two cultures and languages but there is no reason to believe that this is true. A symptom-item could be an excellent discriminator between high and low levels of depression in one culture/language but may be a terrible discriminator in another culture/language. Thus using this item in computing a severity score may provide a biased estimate of the underlying mental health disorder of interest (e.g. major depressive disorder).

Classical Test Theory versus Item Response Theory

The majority of mental health measurement is based on classical test theory (CTT). The test score is a simple summation of the individual item responses, each rated on an ordinal scale by either a clinician or self-reported (Patient Reported Outcome — PRO). Measurement is a counting operation and the outcome of which is dependent on the presentation of the same items to each individual. Item Response Theory (IRT), on the other hand, takes a model-based approach to measurement. In the present context, there is a statistical model which relates characteristics of the items to the severity of illness of the patient. Its origins date back to the pioneering work of Lawley (1943) and Lord (1952). In IRT, the observed responses (e.g. symptom severity ratings) arise from underlying quantitative variation in a latent variable of interest (e.g. depression) which is discretized by an intervening threshold process. The corresponding probability of a specified response to an item (or categorical rating of a symptom by either the patient or the clinician) is a function of the underlying severity of the illness of the patient and characteristics of the items, both of which can be estimated statistically from the binary or ordinal response patterns.

Most applications of IRT are based on the assumption that the latent variable of interest (e.g. mathematical ability) is unidimensional. However, mental health constructs are inherently multidimensional, where items are selected from subdomains which fully characterize the disorder of interest. For example, depression is a multidimensional construct with items drawn from subdomains which include mood disorder, somatization, cognitive impairment, and suicidality. Bock and Aitkin (1981) provided the first multidimensional extension of IRT with an efficient method of estimation. Gibbons and Hedeker (1992) developed the first confirmatory multidimensional IRT model based on the bifactor restriction which allows each item to load on a primary dimension of interest (e.g. depression) and the subdomain dimension from which it was drawn (e.g. somatization). The bifactor model provides a solution which preserves the core dimension of interest, but permits residual correlation among the items within each of an unlimited number of subdomains. It has the advantage of being rotationally invariant and computationally tractable regardless of the number of dimensions, neither of which is true for the general unrestricted multidimensional IRT model. Furthermore, Gibbons and colleagues (2012) have shown that it is directly applicable to computerized adaptive testing, where the interest is in scoring the primary dimension (e.g. depression), but the scale is inherently multidimensional.

Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) makes use of the property of scaled measurement inherent in IRT: different subjects can respond to different items yet be similarly measured in terms of the latent attribute of interest (Weiss, 1985). CAT requires that a large bank of items (i.e. hundreds) be previously calibrated using an IRT model so that those items that are good discriminators of high and low levels of the characteristic of interest can be identified and ordered according to their estimated severity on the latent variable metric of interest (e.g. depression). After each item is administered, an estimate of the patient’s severity is estimated along with its uncertainty. Based on that severity estimate, the next most informative item is then selected and administered based on a statistical selection criterion.

The process continues until a predefined uncertainty threshold (e.g. 5 points on a 100 point scale) is met. Thus, the fundamental difference between CTT and IRT-based CAT is the difference between fixed-length tests with scores of varying precision (CTT) versus variable-length tests with scores of fixed precision (IRT-based CAT). CAT has recently been extended to incorporate the inherent multidimensionality of mental health constructs (Gibbons et al., 2012, 2013, 2014). By analogy, if we created and calibrated a 1000 item bank of mathematics items ranging in difficulty from arithmetic to calculus and had two examinees, one in the 4th grade and another in college, CAT would begin by administering an algebra item and when the 4th grader got it wrong would move to easier items, but when the college student got it right it would move to more difficult items. How far to move and exactly which item to administer next is the statistical key to the problem, which is of course far more difficult for multidimensional constructs such as depression, anxiety, or mania than it is for essentially unidimensional constructs like mathematics. The bifactor model (Gibbons & Hedeker, 1992) is a multidimensional IRT model that is particularly well suited to multidimensional CAT (Gibbons et.al. 2012; Gibbons et.al. 2016).

While it is a relatively easy matter to construct a test based on CTT (select the items and score a subject based on the sum of the manifest item responses), the same is not true for CAT-based IRT. Here we must begin by (a) constructing a large “item bank” consisting of hundreds of items, (b) administer the item bank either in parts or in entirety to a sample of subjects, (c) calibrate the entire item bank using an IRT model, (d) simulate adaptive testing from the complete item responses and select the final CAT tuning parameters, and (e) validate the CAT test scores in a new sample. Once these steps are completed; however, the CAT is much easier to administer and score, is faster, produces test scores with higher precision and accuracy and is more flexible than traditional CTT-based instruments.

The CAT-Mental Health

The CAT-Mental Health (CAT-MH) (Gibbons et.al. 2012, 2013, 2014) is a suite of multidimensional IRT-based CATs for the dimensional measurement of depression, anxiety, mania/hypomania and suicidality. They are based on an item bank of over 1000 items but can reproduce the full bank test scores with high correlation (r=0.95) based on adaptive administration of an average of 36 items (approximately 12 for each of the three constructs) in an average of approximately 6 minutes (2 minutes each) via the internet on any internet capable device. Since the same items are not repeatedly administered, response bias associated with traditional fixed-length measures such as the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al. 2001) is eliminated. Nevertheless, test-retest reliability has been shown to be higher for CAT (r=0.92) than for traditional fixed length depression test scores such as the PHQ-9 (r=0.84) (Beiser, Vu, & Gibbons, 2016).

Differential Item Functioning

Traditional psychometric instruments are adapted to a different language and culture through a careful forward and backward translation followed by adjudication of unresolved differences by a multinational committee. However, even the best translation does not insure that the meaning of the items and their ability to discriminate high and low levels of the underlying construct (e.g. depression) are the same in different populations (e.g. Latinos versus non-Latinos).

Differential item functioning (DIF) occurs when individuals at the same level on the trait(s) or construct(s) being assessed but from different subpopulations have unequal probabilities of attaining a given score on the item. Methods for investigating DIF have been developed for both dichotomously and polytomously scored items. These methods may be classified by whether they condition on an unobserved or observed variable. Item response theory, logistic regression, and Mantel-Haenszel procedures for dichotomously scored responses and their extensions to polytomous responses are currently the most widely used methods for detecting DIF. The IRT approach (Thissen et.al. 1993; Cai et.al. 2011) usually involves the comparison of two unidimensional models, a compact model (with common parameters between the different subpopulations) and an augmented model where a subset of the parameters are allowed to vary across the subgroups. Glas (Glas, 1998; Glas 1999) argued that the LR and Wald test approach to DIF detection are not efficient because they require estimation of the parameters of IRT model under the alternative hypothesis of DIF for every single item. Therefore, Glas proposed using the Lagrange multiplier test, which does not require estimation of the parameters of the alternative model.

Kim et.al, (2016) introduced a new approach to estimating DIF that is suitable for both multidimensional IRT (based on a bifactor model) and CAT. The idea is based on fitting a logistic regression model for each item’s observed response category on the estimated severity score for the primary dimension of interest based on the original calibration, which can be based on any unidimensional or multidimensional (bifactor) IRT model. Items or symptoms exhibiting DIF will have a weaker or stronger relationship to the estimated test (severity) score for the primary dimension of the bifactor model based on the original calibration. In this way, items that do not discriminate high versus low levels of severity in the target population can be easily identified. This method permits DIF testing on item responses obtained from routine CAT administration. The major advance is that testing DIF in a new population (e.g. Latinos) can be performed by simply administering the existing CAT to a representative sample drawn from the target population, insuring that the scale is valid in a new population. The ultimate goal is to identify poorly discriminating items for Latinos taking these tests in Spanish so that they can be eliminated from the adaptive tests when administered in Spanish, improving the quality of measurement in this culture. Finally, this approach incorporates the multidimensionality of the mental health constructs directly into the DIF computation. By contrast, application of traditional DIF based on unidimensional IRT may find evidence of DIF that is produced by failure to account for the other dimensions.

Method

Participants

Prior to the enrollment of any participants, the study received research ethics committee approval from the Institutional Review Board after it was determined to pose no more than minimal risk to study participants.

Non-Latino sample.—The original CAT-MH was developed using an initial calibration sample of 798 male and female adult (18–80 years of age) psychiatric outpatients from the Western Psychiatric Institute and Clinic (WPIC) at the University of Pittsburgh and a community clinic at DuBois Regional Medical Center (DuBois RMC), see Table 1. Patients were excluded if they had a confirmed (medical records and treating physician) diagnosis of schizophrenia, schizoaffective disorder, or psychosis; organic neuropsychiatric syndromes (e.g., Alzheimer disease); drug or alcohol dependence within the past 3 months (patients with episodic abuse related to mood episodes were not excluded); inpatient treatment status; and inability to provide informed consent (see Gibbons et al., 2012).

Latino sample.—Spain has the largest number of Latino migrants in the European Union (Connor & Massey, 2010), only surpassed by the U.S., where most immigrants are Latino (Acosta & de la Cruz, 2011; Padilla & Peixoto, 2007). Migrants comprise 12.2% of Spain's population, with close to a third (28%) being Latinos from the Caribbean, Central and South America (Connor & Massey, 2010). Migrants in the U.S. represent 15% of the U.S. population and 53% of these are Latino (U.S. Department of Homeland Security, 2011). We recruited 1,276 patients (422 from Boston, 354 from Madrid and 500 from Barcelona) through direct contact in clinic waiting rooms from mental health, substance use, primary care, and HIV clinics in Boston, Massachusetts, and Madrid and Barcelona, Spain (see Table 1). The Latino populations in both Spain and Massachusetts were diverse. First-generation Latino immigrants (born in a country other than the interview site) comprised 99.5% of the Spain sample and 95.0% of the Massachusetts sample. Massachusetts participants were born in Central America (49.8%), Puerto Rico (18.5%), the Caribbean (12.6%), South America (13.5%), and Continental U.S. (5.0%). The majority of participants in Spain were of South American origin (80.2% in Madrid and 86.8% in Barcelona), followed by the Caribbean (12.7%) in Madrid and Central American (7.6%) in Barcelona. Extensive training in the research protocol was provided even with interviewers who shared language to address variations in interviewers' familiarity with research methodology and prevent incorrect interpretations of concepts. To accomplish this, the research teams in Boston, Madrid and Barcelona, and colleagues from Puerto Rico (experts in translation and adaptation of instruments) worked on translating and adapting the CAT-MH to Spanish using well-known methodology (Matías-Carrelo et al., 2003) to attain a Spanish version with semantic, content and technical equivalence to the original English version. First the second author sent to a professional translation company to translate all the CAT-MH items from English to Spanish. A separate team of bilingual investigators (LC and two research assistants) reviewed the translation and identified that some of the terms were not identical to the English terms in describing the symptoms or questions. Two research investigators (MA and GC) that were fully bilingual and had worked on translations of diagnostic and symptoms measures revised the professional translation of the 1008 items to change some literal translations to ensure better content equivalence. After completing this step, the modified Spanish translation was sent to two bilingual investigators to translate all the 1008 CAT-MH items from Spanish to English (LH, SM). The four investigators of which two have also been clinicians (MA, GC, SM, and LH) then held a conference call to determine how the English back translation differed from the original English version. In those items were there were differences between the back translation from Spanish to English and the original English version, the team of investigators reviewed the discrepancies and determined how to make sure that the Spanish item was consistent with the English version. We then set up a multinational bilingual committee (MNBC) formed by four researchers and four clinicians that included Spanish speakers from six diverse countries or territories (Puerto Rico, Mexico, Panama, Columbia, Spain, Peru) to ensure that the language selected would be the most neutral possible across Spanish-speaking countries. The MNBC was also convened to determine if the translated and adapted Spanish items that had been back translated from English to Spanish had both semantic and conceptual equivalence to the English items.

To accomplish this task, three conference calls were held to discuss the items where individually the committee members had the identified problems with the Spanish translation or where items needed to be reworded to be more conversational or colloquial in Spain or in Boston. MNBC members individually sent their written suggestions for changes to the items before the call and these were discussed in the conference call. After discussion, the wording of items preferred by the MNBC was circulated and sometimes discussed again, until there was agreement between the members. The MNBC determined final decisions for each item through unanimous consensus between the members after this process. The CAT-MH was piloted with a small group of Latino migrant workers to see if the wording was understood and to ensure necessary adjustments, with some items modified from their original version to ensure that respondents with lower education understood the wording of items. We used cognitive debriefing and asked these respondents to identify items that were difficult to answer or confusing. Based on their responses, we either change some words or change the sequence of words to better capture the meaning of the item. In all cases the items remain comparable, if not identical to the original items. All 1008 items in the bank were translated and adapted so that adaptive administration of the entire CAT-MH could ultimately be conducted in Spanish.

Recruitment began in March of 2014 and final screening interviews were conducted in August of 2015. Participants came from community-based clinics and organizations serving a diverse population of Latinos, such as Latino social services organizations and substance use treatment facilities, many of whom serve a low income, safety-net population. We trained bilingual research staff to conduct the research protocol using interviewing techniques, and gave them opportunities to conduct mock interviews before entering the field. To ensure quality control, a quality check of the first 3 cases of all new interviewers and a randomized 15% of additional cases was instituted. Quality control included listening to the audio recordings and filling out a checklist of critical areas to ensure interview quality and accuracy. After they were certified to conduct the protocol, they recruited patients in waiting areas of clinics, by referral in community groups, or by telephone screening of patients that had consented to be told about the study. Approval for the study was obtained from the institutional review boards by all participating institutions.

Item Bank

The item bank has been described in detail by Gibbons et.al. (2012). Symptomatology was evaluated during the past 2 weeks based on patient self-reports. The bank included a total of 1008 items distributed across the three primary domains of depression (452 items), anxiety

(437 items) and mania/hypomania (89 items). The items were selected based on review of over 100 existing depression or related rating scales and were modified to have a limited set of consistent response categories. Most items were rated on a 5-point (Likert) ordinal scale. Example items are provided in the on-line supplement of the previously published paper (Gibbons et al., 2012).

Design.—As previously described, the entire the entire 1008 item bank was translated and reverse translated into Spanish and adjudicated by a multinational bilingual expert committee for conceptual and linguistic equivalence. This was done so that the entire CAT-MH could be administered in Spanish for routine assessment. However, in practice only approximately 200 items are commonly used for adaptive testing and for any given person, an average of only 36 items are administered to assess depression, anxiety and mania, and these items are drawn from each of the relevant subdomains from each of the three CAT-MH scales. To examine the less frequently administered items, we used a balanced incomplete block (BIB) design (Gibbons et al., 2012), and created 36 forms of 25 items each that were drawn from items that were not administered as a part of the adaptive tests. As such, on average, each subject received 36+25=61 items. Our DIF analysis is based on 1276 participants who took the Spanish version of the CAT-MH via tablet computer and also received an additional 25 items drawn from the less commonly administered items. The enormous costs and degree of subject burden associated with administration of the full 1008 item bank made this unfeasible, and led us to this continuous quality improvement design, which focuses on the key items used in adaptive testing of these constructs.

Statistical Methods

The Latino sample patients were given CAT administration of depression, anxiety and mania tests and scored using the methodology originally described by Gibbons and colleagues (Gibbons et al., 2012, 2016). We used the new method for assessing DIF designed specifically for CAT-based administration of multidimensional tests (Kim et al., 2016). The model examines DIF in terms of the primary severity dimension based on a bifactor model (Gibbons and Hedeker, 1992; Gibbons et.al, 2007). The general strategy involved the following algorithm:

- The original non-Latino psychiatric sample calibration was used to score the Latino subject's item responses for each of the three constructs (depression, anxiety, and mania/hypomania).
- For each item, an ordinal logistic regression was used to model the association between the estimated test score based on the non-Latino calibration and the actual response category selected by each of the Latino subjects.
- To incorporate uncertainty in the test score, the process was repeated 100 times drawing new values from the posterior distribution of the test score.
- Kim et.al. (2016) suggest that values less than 1.0 for the logistic regression slope indicate DIF. This is equivalent to an odds ratio of exp(1.0)=2.72, or a 2.72 increase in the likelihood of changing one response category for each unit increase in the underlying severity score or approximately a 15-fold increase

across the range of the scale scores (the scale scores are measured on an underlying unit normal scale (N(0,1)) which has range from −3 to +3).

As a final test of DIF, we also fit a new bifactor model to the Latino data and compared the estimated primary dimension test scores for each domain from that model to the original non-Latino calibration based test score, before and after eliminating those items exhibiting DIF. This allows us to directly examine the extent to which the optimal calibration for the Latino data produces severity estimates which differ from those based on the original calibration. Factor loadings for the original and target populations can then be examined to determine the magnitude and direction of DIF.

The original item bank contained 1008 items, and it was not feasible to administer the entire item bank to the Latino sample. Instead, we tested the most commonly administered items (based on CAT administration and an optimally selected set of 25 additional items for each subject) for DIF. Not all of the potentially adaptively administered items are included in the final DIF testing, but the most frequently administered items are. These items had a minimum of 80 participants responding to the item. There were 81 depression, 84 anxiety and 59 mania items used in the DIF analyses. In addition to testing for DIF, we computed the percentage of patients screening positive for MDD using the brief (4–6 items) computerized adaptive diagnostic screener CAD-MDD (Gibbons, et.al. 2013).

Finally, we examined the fit of the bifactor model to each of the three domains in the Latino sample. Traditional limited information fit statistics such as RMSEA, CFI and TL, used in structural equation models (SEM) do not directly apply to categorical item-response data and full-information maximum marginal likelihood estimation used in fitting the bifactor model. For full-information models, we use likelihood ratio chi-square statistics to examine the improvement in fit of a bifactor model (or any multidimensional model) over a unidimensional alternative. This has been done in all of the original publications and provides evidence of significant improvement in model fit of the bifactor model over a traditional unidimensional alternative (Gibbons et.al. 2012). This is also repeated here for the Latino sample for each of the three domains. Assessment of absolute fit for full-information MMLE is complicated by the fact that there are k^n possible response patterns (where k is the number of categories and n is the number of items), and in most samples, there is typically a single realization of the subset of the response patterns observed in that study. To this end, we have presented figures which show the item-category level comparison of the marginal estimated and observed frequencies. High correlation between the observed and expected item-category level proportions indicates excellent absolute fit of the model to the observed data.

Results

Demographic Distributions

Table 1 presents the demographic distributions of age, gender, ethnicity, race and education in the Latino and non-Latino samples. The largest difference is ethnicity with 100% versus 1% Latino. Gender is well balanced between the two cohorts. The Latino respondents were younger and of a lower educational level than non-Latinos. In addition, the Latino patients had a much higher designation of “mixed race” than the non-Latino sample.

Rates of MDD and Severity

In the Latino sample, the rate of MDD based on the CAD-MDD (with 90% confidence level) was 25%. Among those with MDD, 9% were in the moderate or severe categories (scores of 65 or greater on a 100 point scale) based on their CAT-estimated score.

DIF

Table 2 presents a listing in English and Spanish of the 8 items that exhibited significant DIF.

Depression DIF.—We identified the following 4 items with DIF that failed to discriminate between high and low levels of depression in the Latino sample. Referring to the past two weeks:

1. How much have you felt cheerful?
2. I felt satisfied.
3. How easily did you get tired?
4. I had trouble keeping my mind on what I was doing.

The correlation between the depression severity scores based on the original calibration and the new Latino calibration was r=0.985 for all items and r=0.990 eliminating these 4 items. Figure 1a shows the bivariate relationship between the estimated depression severity scores based on the original non-Latino calibration and the new Latino calibration after eliminating the 4 poor discriminating items. The Figure graphically illustrates the high correlation reported above and verifies that the item parameters from the non-Latino sample and Latino sample produce virtually identical depression test scores. The direction of the DIF was for lower discrimination in the Latino sample for all 4 items. Estimated primary factor loadings were 0.74 vs. 0.32 for “felt cheerful”; 0.78 vs. 0.28 for “felt satisfied”; 0.84 vs. 0.51 for “get tired easily”; and 0.79 vs. 0.47 for “keeping my mind on what I was doing,” for non-Latino vs. Latino samples respectively. In the non-Latino sample the range of factor loadings on the primary depressive severity dimension was 0.67 to 0.91. For the Latino sample, the range of factor loadings for items not exhibiting DIF was 0.51 to 0.97. The bifactor model provided significant improvement in fit over a unidimensional IRT model (chi-square=420.14, df=81, p<0.0001). Excellent fit of the bifactor model to the Latino data is illustrated in Figure 2a, which displays the observed and estimated marginal response category proportions for which the correlation is r=0.952.

Anxiety DIF.—There were no anxiety items that exhibited DIF. The correlation between the anxiety severity scores based on the original calibration and the new Latino calibration was r=0.993 (see Figure 1a). In the non-Latino sample the range of factor loadings on the primary anxiety severity dimension was 0.65 to 0.89. For the Latino sample, the range of factor loadings was 0.60 to 0.90. The bifactor model provided significant improvement in fit over a unidimensional IRT model (chi-square=715.20, df=84, p<0.0001). Excellent fit of the bifactor model to the Latino data is illustrated in Figure 2b, which displays the observed and estimated marginal response category proportions for which the correlation is r=0.960.

Mania DIF.—We identified the following 4 items with DIF that failed to discriminate between high and low levels of mania in the Latino sample. Referring to the past two weeks:

1. Did you ever engage in risk-taking behaviors, such as driving fast, promiscuous sex, hanging out in dangerous neighborhoods?
2. Have you had periods of at least 3 days in which you were preoccupied with yourself and your own problems, thoughts, and feelings?
3. Have you had periods of at least 3 days in which you felt self-assured, charismatic or tended to assume a leadership role?
4. Have you had periods of at least 3 days in which you were less sexually active than is typical for you?

The correlation between the mania severity scores based on the original calibration and the new Latino calibration was r=0.962 (see Figure 1c). The direction of the DIF was for lower discrimination in the Latino sample for the 1st two items and higher discrimination for the latter two items. Estimated primary factor loadings were 0.53 vs. 0.18 for “risk taking behaviors”; 0.63 vs. 0.57 for “preoccupied with self”; 0.34 vs. 0.42 for “self-assured and charismatic”; and 0.43 vs. 0.63 for “less sexually active,” for non-Latino vs. Latino samples respectively. Despite the differences in direction of factor loadings, none of these 4 items met the criteria for good discrimination and should therefore not be included for routine practice in Latinos taking these tests in Spanish. In the non-Latino sample the range of factor loadings for items not exhibiting DIF on the primary mania severity dimension was 0.37 to 0.79. For the Latino sample, the range of factor loadings for items not exhibiting DIF was 0.30 to 0.97. The bifactor model provided significant improvement in fit over a unidimensional IRT model (chi-square=250.08, df=59, p<0.0001). Excellent fit of the bifactor model to the Latino data is illustrated in Figure 2c, which displays the observed and estimated marginal response category proportions for which the correlation is r=0.953.

Discussion

Overall there was very little evidence of DIF in Latinos taking the CAT-MH in Spanish relative to the original non-Latino calibration sample assessed in English. For all three tests there were a total of only 8 items that exhibited significant DIF. This was an unexpected finding. Overall, correlations between test scores based on our original calibration sample taking the test in English and the Latino sample taking the test in Spanish were extremely high (depression and anxiety tests were both r=0.99 and r=0.96 for mania). This result provides confidence that differences between severity levels measured by the CAT-MH between Latino and non-Latino populations are real differences and not an artifact of cultural bias in the measurement process itself. This largely null finding is important because it provides confidence that in general the experience of depression, anxiety, and mania are similar in these two cultural groups taking these adaptive tests in their native languages and that the meaning of the items is not culturally dependent. The CAT-MH can therefore be reliably used in its slightly modified form (i.e. eliminating the 8 items with DIF) to assess the severity of depression, anxiety and mania in Latinos taking these adaptive tests in Spanish.

There are potential explanations for why these depression items (cheerfulness, satisfied with life, easily get tired, and trouble keeping my mind on what I was doing) have DIF. The concept of cheerfulness is not easily translated to Spanish, where similar words like animated, jovial, or happy could be used but are not identical in the degree of positive feeling. “Animated” was selected by the Multinational Bilingual Committee but might not represent the level of joyfulness embodied in cheerfulness. “I felt satisfied” is mostly an assertion used in Spanish after eating a meal, and less a concept rating one’s life. “Easily get tired” or “having trouble keeping my mind on what I was doing” are unlikely to differentiate Latino workers with and without depression that have two or three jobs and are logically tired or are in professions like manufacturing or service cleaning, where operating on automatic pilot is necessary to survive the monotony of the job.

The reasons for DIF items for Mania are more complicated. The notion of “risk-taking behaviors” is relative depending on your life circumstances (i.e. gang violence or unsafe neighborhoods), with many Latinos living in dangerous neighborhoods because of poverty or driving fast to avoid getting killed, actions that might not represent mania symptoms. In Latino culture, a collectivistic society, the view that “you were preoccupied with yourself and your own problems, thoughts, and feelings” is seen as pejorative and selfish, and less likely to be endorsed due to social desirability. The concepts of self-assurance, charisma and leadership roles are less likely to be characteristically used as adjectives by Latinos, especially when experiencing marginalization and oppression both by poverty and minority status. “Being less sexually active than is typical for you”, might be related to having being separated from ones sexual partners or loved ones as part of immigration, and therefore, less likely to be a good predictor of mania. As discussed, Latinos’ life experiences can influence the interpretation of research questions in unique ways.

While the largest difference between the two samples was Latino ethnicity, and the language used to administer the test, there other sociodemographic differences worth noting that could also lead to DIF. The Latino sample was somewhat younger and had a lower educational level. Differences in racial composition were largely related to the Latino patients having a much larger designation of “mixed-race” than the non-Latino patient sample. The general lack of DIF; however, suggests that these imbalances are not having a large effect on the understanding of the items or the experience of these three mental health disorders.

The logistic regression approach to estimating DIF used here has several advantages over traditional approaches. The method preserves the multidimensional nature of the IRT model used in the calibration and can be used effectively in much smaller sample sizes since it is applied individually to each item. As such it is similar to traditional methods for DIF analyses such as the Mantel-Haenszel test (Holland & Wainer, 1993) with the improvement that it relies upon more sophisticated estimates of severity scores for the target group based on the original calibration data and can accommodate both unidimensional and multidimensional IRT model parameterizations. Furthermore, the use of a continuous quality improvement model in which the adaptive versions of the tests are administered in the target population (and a subset of items less frequently administered) and DIF is assessed in these items, provides an excellent alternative to full item-bank administration in the target population which is often not possible. This makes it possible to test DIF in numerous cultures, languages, demographics, and indications, based on routine application of the CATs in the target population of interest.

In summary, we found that only a handful of items (8 out of 124) that are routinely adaptively administered via the CAT-MH (supplemented by a selection of 25 additional items from the 1008 item bank for each subject) exhibited DIF in Latinos taking the tests in Spanish. Changing the scoring of the test to accommodate these differences had negligible effects; however, a conservative solution is to eliminate these items from test administrations in Spanish. This is the enormous advantage of basing CAT on large item banks. The addition or deletion of a few items will in no way bias the scoring of the underlying dimensions of interest and restricting items to those with good discrimination will insure that our measurements are of high and consistent quality between different cultures and/or in patients with different comorbidities.

Future DIF studies should consider independent replication of the results of this study as well as extensions of the methodology to individual subethnicities (South American, Puerto Rican, Mexican, and Cuban) as larger sample sizes become available with sufficient representation of these different groups. Given the inclusion of multiple subethnicities in the current sample and the general absence of DIF identified, it is unlikely that nativity-specific forms of DIF will emerge. Nevertheless, this is an empirical question and should be studied further.

Acknowledgments

This study was supported by an administrative supplement from the National Institutes of Health (NIH) for A New Statistical Paradigm for Measuring Psychopathology Dimensions in Youth (Grant #: 3R01MH100155-01S1) and by the National Institute On Drug Abuse (NIDA) of NIH under Award Number R01DA034952

References

Acosta YD, de la Cruz GP (2011). The Foreign Born from Latin America and the Caribbean: 2010. (Report No.: ACSBR/10–15), Washington DC: U.S. Department of Commerce: Economics and Statistics Administration, 6pp.

Aguilera A, Garza MJ, Muñoz RF (2010). Group cognitive-behavioral therapy for depression in Spanish: culture-sensitive manualized treatment in practice. Journal of Clinical Psychology, 66(8), 857–867. [PubMed: 20549680]

Akincigil A, Olfson M, Siegel MJ, Zurlo KA, Walkup JT, Amin S, Crystal S (2012). Racial and ethnic disparities in depression care in community dwelling elderly in the United States. American Journal of Public Health, 102(2), 319–328. [PubMed: 22390446]

Alarcon RD, Bell CC, Kiermayer LJ, Lin KM, Ustun B, Weisner KL (2002). Beyond the funhouse mirrors: Research agenda on culture and psychiatric diagnosis In: Kupfer D, First B, Regier D, (Eds.). A Research Agenda for DSM-IV, Arlington, VA: American Psychiatric Association.

Alegria M, Canino G, Rios R, Vera M, Calderon J, Rusch D, Ortega AN (2002). Inequalities in use of specialty mental health services among Latinos, African Americans, and non-Latino whites. Psychiatric Services, 53(12), 1547–1555. [PubMed: 12461214]

Alegria M, Canino G, Shrout PE, Woo M, Duan N, Vila D, Meng XL (2008). Prevalence of mental illness in immigrant and non-immigrant U.S. Latino groups. The American Journal of Psychiatry, 165(3), 359–369. [PubMed: 18245178]

Alegria M, Mulvaney-Day N, Torres M, Polo A, Cao Z, Canino G (2007). Prevalence of psychiatric disorders across Latino subgroups in the United States. American Journal of Public Health, 97(1), 68–75. [PubMed: 17138910]

Angel J, Angel R (2006). Minority group status and healthful aging: Social structure still matters. American Journal of Public Health, 96(7), 1152–1159. [PubMed: 16735614]

Angel JL, Whitfield KE (2007) Setting the Stage: Hispanic Health and Aging in the Americas In: Angel JL, Whitfield KE (Eds.). The Health of Aging Hispanics: The Mexican-Origin Population. New York, NY: Springer.

Beiser D, Vu M, Gibbons RD (2016). Test-retest reliability of a computerized adaptive depression screener. Psychiatric Services, 67(9), 1039–1041. [PubMed: 27079989]

Bock RD, Aitkin M (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443–459.

Breslau J, Aguilar-Gaxiola S, Kendler KS, Su M, Williams D, Kessler RC (2006). Specifying race-ethnic differences in risk for psychiatric disorder in a USA national sample. Psychological Medicine, 36, 57–68. [PubMed: 16202191]

Breslau J, Javaras KN, Blacker D, Murphy JM, Normand S-LT (2008). Differential item functioning between ethnic groups in the epidemiological assessment of depression. The Journal of Nervous and Mental Disease, 196(4), 297. [PubMed: 18414124]

Cai L, Yang JS, Hansen M (2011). Generalized full-information item bifactor analysis. Psychological Methods, 16(3), 221–248. [PubMed: 21534682]

Cauce AM, Domenech-Rodriguez M, Paradise M, Cochran BN, Shea JM, Srebnik D, Baydar N (2002). Cultural and contextual influences in mental health help seeking: A focus on ethnic minority youth. Journal of Consulting and Clinical Psychology, 70(1), 44–55. [PubMed: 11860055]

Chatterji P, Alegria M, Takeuchi D (2009). Racial/ethnic differences in the effects of psychiatric disorders on employment. Atlantic Economic Journal, 37(3), 243–257. [PubMed: 19898677]

Choi H (2002). Understanding adolescent depression in ethnocultural context. Advances in Nursing Science, 25(2), 71–85. [PubMed: 12484642]

Compton WM, Conway KP, Stinson FS, Colliver JD, Grant BF (2005). Prevalence, correlates, and comorbidity of DSM-IV antisocial personality syndromes and alcohol and specific drug use disorders in the United States: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. The Journal of Clinical Psychiatry, 66(6), 677–685. [PubMed: 15960559]

Connor P, Massey DS (2010). Economic outcomes among Latino migrants to Spain and the United States: Differences by source region and legal status. The International Migration Review, 44(4), 802–829. [PubMed: 21776179]

Cooper LA, Gonzales JJ, Gallo JJ, Rost KM, Meredith LS, Rubenstein LV, Wang NY, Ford DE (2003). The acceptability of treatment for depression among African-American, Hispanic, and white primary care patients. Medical Care, 41(4): 479–489. [PubMed: 12665712]

Crockett LJ, Randall BA, Shen Y-L, Russell ST, Driscoll AK (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. Journal of Consulting and Clinical Psychology, 73(1), 47. [PubMed: 15709831]

D'Angelo E, Llerena-Quinn R, Shapiro R, Colon F, Rodriguez P, Gallagher K, Beardslee W (2009). Adaptation of the preventive intervention program for depression for use with predominantly low-income Latino families. Family Process, 48(2): 269–291. [PubMed: 19579909]

deGruy FV, Pincus HA (1996). The DSM-IV-PC: a manual for diagnosing mental disorders in the primary care setting. The Journal of the American Board of Family Practice, 9(4), 274–281. [PubMed: 8829077]

Dunlop DD, Manheim LM, Song J, Lyons JS, Chang RW (2005). Incidence of disability among Preretirement Adults: The impact of depression. American Journal of Public Health, 95(11), 2003–2008. [PubMed: 16254232]

Favazza A, Oman M (1984). Overview: Foundations of cultural psychiatry In: Mezzich JE, Berganza CE (editors). Culture and Psychopathology (pp 17–31). New York, NY: Columbia University Press.

Gibbons RD, Bock RD, Hedeker D, Weiss D, Segawa E, Bhaumik DK, Kupfer D, Frank E, Grochocinski V, Stover A (2007). Full-Information Item Bi-Factor Analysis of Graded Response Data. Applied Psychological Measurement, 31, 4–19.

Gibbons R, Hedeker D (1992). Full information item bi-factor analysis. Psychometrika, 57(3), 423–436.

Gibbons RD, Hooker G, Finkelman MD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kupfer DJ (2013). The CAD-MDD: A computerized adaptive diagnostic screening tool for depression. Journal of Clinical Psychiatry, 74(4), 669–674. [PubMed: 23945443]

Gibbons R, Weiss DJ, Pilkonis P, Frank E, Moore T, Kim JB, Kupfer D (2012). Development of a computerized adaptive test for depression. Archives of General Psychiatry, 69, 1104–1112. [PubMed: 23117634]

Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, Kupfer DJ (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. The American Journal of Psychiatry, 171(2), 187–194. [PubMed: 23929270]

Gibbons RD, Weiss DJ, Frank E, Kupfer DJ (2016). Computerized adaptive diagnosis and Testing of mental health disorders. The Annual Review of Clinical Psychology. 12, 83–104.

Glas CAW (1998). Detection of differential item functioning using Lagrange multiplier tests. Statistica Sinica, 8(3), 647–668.

Glas CA (1999). Modification indices for the 2-PL and nominal response model. Psychometrika, 64(3), 273–294.

Holland P, Wainer H (Eds.) (1993). Differential Item Functioning. Hillside, NJ: Lawrence Erlbaum Associates.

Hovey JD (2000). Acculturative stress, depression, and suicidal ideation in Mexican immigrants. Cultural Diversity & Ethnic Minority Psychology, 6(2), 134–151. [PubMed: 10910528]

Hovey JD (2000). Acculturative stress, depression, and suicidal ideation among Central American immigrants. Suicide & Life-Threatening Behavior, 30(2), 125–139. [PubMed: 10888053]

Huang FY, Chung H, Kroenke K, Spitzer RL (2006). Racial and ethnic differences in the relationship between depression severity and functional status. Psychiatric Services, 57(4), 498–503. [PubMed: 16603745]

Kim JJ, Silver RK, Elue R, Adams MG, La Porte LM, Cai L, Kim JB, Gibbons RD (2016). The experience of depression, anxiety and mania among perinatal women. Archives of Women's Mental Health, 19(5), 883–890.

Kleinman A (1988). Rethinking Psychiatry: From Cultural Category to Personal Experience. New York, NY: Free Press.

Kroenke K, Spitzer RL, Williams JB. (2001). The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med, 16, 606–613. [PubMed: 11556941]

Lagomasino IT, Dwight-Johnson M, Miranda J, Zhang L, Liao D, Duan N, Wells K (2005). Disparities in depression treatment for Latinos and site of care. Psychiatric Services, 56(12), 1517–1523. [PubMed: 16339612]

Lawley DN (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61, 273–287.

Le HN, Perry DF, Stuart EA (2011). Evaluating a preventive intervention for perinatal depression in high-risk Latinas. Journal of Consulting and Clinical Psychology, 79(2): 135–141. [PubMed: 21319897]

Lewis-Fernández R, Hinton DE, Laria AJ, Patterson EH, Hofmann SG, Craske MG, Stein DJ, Asnaani A, Liao B (2010). Culture and the anxiety disorders: recommendations for DSM-V. Depression and Anxiety. 27(2), 212–229. [PubMed: 20037918]

Lord FM (1952). A Theory of Test Scores. Psychometric Monograph, 7.

Malgady RG, Zayas LH (2001). Cultural and linguistic considerations in psychodiagnosis with Hispanics: The need for an empirically informed process model. Social Work, 46(1), 39–49. [PubMed: 11217492]

Matias-Carrelo LE, Chavez LM, 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.

McGuire TG, Alegria M, Cook BL, Wells KB, Zaslavsky AM (2006). Implementing the Institute of Medicine definition of disparities: An application to mental health care. Health Services Research, 41(5), 1979–2005. [PubMed: 16987312]

McGuire TG, Miranda J (2008). New evidence regarding racial and ethnic disparities in mental health: Policy implications. Health Affairs, 27(2), 393–403. [PubMed: 18332495]

Mechanic D (2002). Removing barriers to care among persons with psychiatric symptoms. Health Affairs, 21(3), 137–147. [PubMed: 12025977]

Mezzich JE, Kirmayer LJ, Kleinman A, Fabrega H, Parron DL, Good BJ, Manson SM (1999). The place of culture in DSM-IV. The Journal of Nervous and Mental Disease, 187(8), 457–464. [PubMed: 10463062]

Miranda J, Cooper LA (2004). Disparities in care for depression among primary care patients. Journal of General Internal Medicine, 19(2), 120–126. [PubMed: 15009791]

Miranda J, Schoenbaum M, Sherbourne C, Duan N, Wells K (2004). Effects of Primary Care Depression Treatment on Minority Patients' Clinical Status and Employment. Archives of General Psychiatry, 61, 827–834. [PubMed: 15289281]

Mojtabai R (2008). Social comparison of distress and mental health help-seeking in the US general population. Social Science & Medicine, 67(12), 1944–1950. [PubMed: 18977062]

Office of the Surgeon General, Center for Mental Health Services. (2001). National Institute of Mental Health Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Ormel J, VonKorff M, Ustun TB, Pini S, Korten A, Oldehinkel T (1994). Common mental disorders and disability across cultures. Results from the WHO Collaborative Study on Psychological Problems in General Health Care. Journal of the American Medical Association, 272(22), 1741–1748. [PubMed: 7966922]

Padilla B, Peixoto J (2007). Latin American Immigration to Southern Europe. Migration Policy Institute; Available from: http://www.migrationinformation.org/Feature/display.cfm?ID=609.

Santiago-Rivera A, Kanter J, Benson G, Derose T, Illes R, Reyes W (2008). Behavioral activation as an alternative treatment approach for Latinos with depression. Psychotherapy, 45(2): 173–185. [PubMed: 22122415]

Schoenbaum M, Miranda J, Sherbourne C, Duan N, Wells K (2004). Cost-effectiveness of interventions for depressed Latinos. The Journal of Mental Health Policy and Economics, 7, 69–76. [PubMed: 15208467]

Schraufnagel TJ, Wagner AW, Miranda J, Roy-Byrne PP (2006). Treating minority patients with depression and anxiety: what does the evidence tell us? General Hospital Psychiatry, 28(1), 27–36. [PubMed: 16377362]

Sussman LK, Robins LN, Earls F (1987). Treatment-seeking for depression by black and white Americans. Social Science & Medicine, 24(3), 187–196. [PubMed: 3824001]

Teresi JA, Ramirez M, Lai J-S, Silver S (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychology Science Quarterly. 50(4), 538. [PubMed: 20165561]

Thissen D, Steinberg L, & Wainer H (1993). Detection of differential item functioning using the parameters of item response models In Holland P, & Wainer H (Eds.), Differential item functioning (pp. 67–100). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

U.S. Department of Homeland Security. (2011). 2010 Yearbook of Immigration Statistics. Washington, DC: Office of Immigration Statistics.

Ustun TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJ (2004) Global burden of depressive disorders in the year 2000. The British Journal of Psychiatry, 184, 386–392. [PubMed: 15123501]

Vacc NA, Juhnke GA (1997). The use of structured clinical interviews for assessment in counseling. Journal of Counseling and Development, 75(6), 470–480.

Vaughn-Coaxum RA, Mair P, Weisz JR (2016). Racial/Ethnic differences in youth depression indicators: An item response theory analysis of symptoms reported by White, Black, Asian, and Latino youths. Clinical Psychological Science,. 4(2), 239–253. [PubMed: 31289695]

Vega W (2001). Latino mental health treatment in the United States In Aguirre-Molina M, Molina CW, Zambrana RE (Eds.), Health issues in the Latino community, New York, NY: Jossey-Bass.

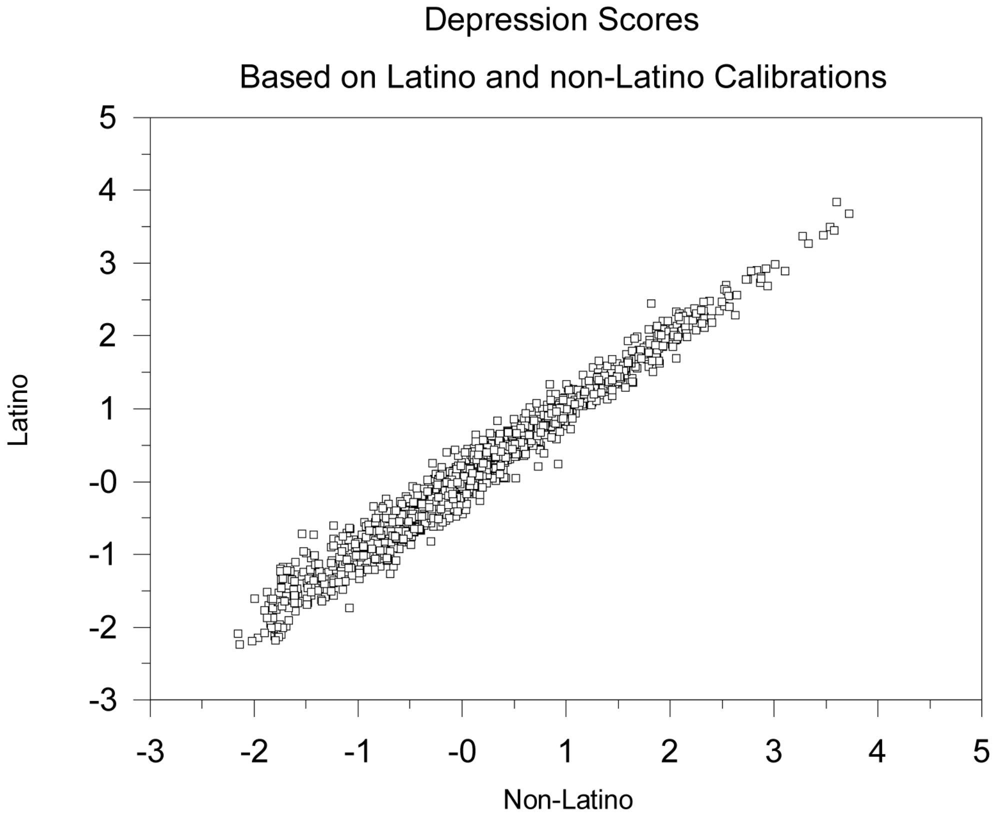
Vega WA, Lopez SR (2001). Priority issues in Latino mental health services research. Mental Health Services Research, 3(4), 189–200. [PubMed: 11859965]

Weiss DJ (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53(6), 774–789. [PubMed: 3841355]

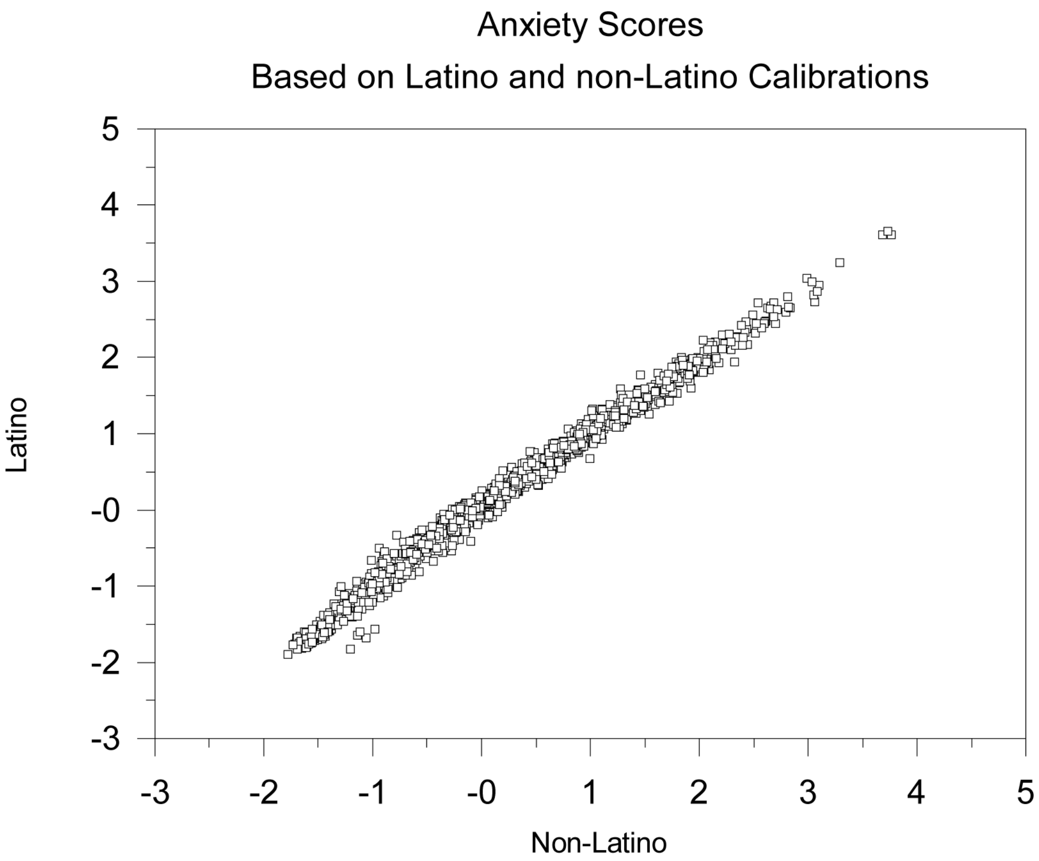
Westermeyer J, Janca A (1997). Language, culture and psychopathology: Conceptual and methodological issues. Transcultural Psychiatry, 34, 291–311.

Public Significance Statement:

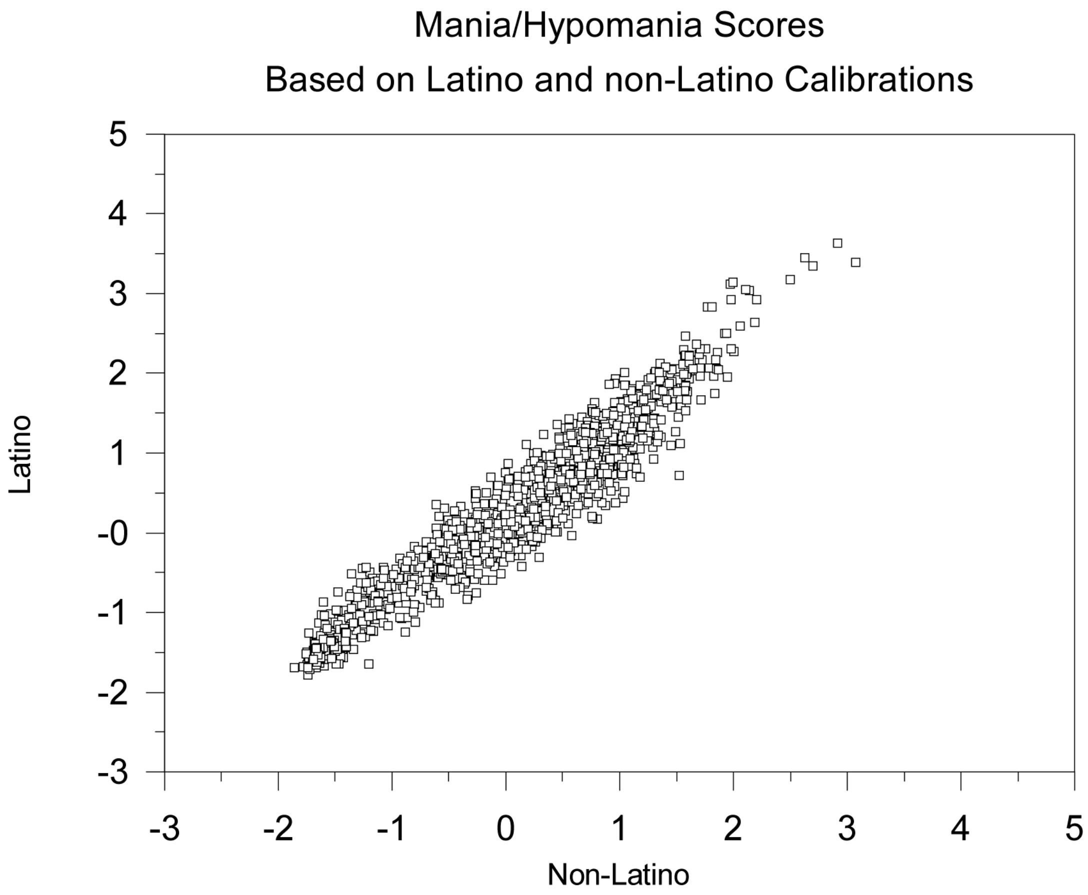
This is the first study to validate the use of computerized adaptive tests for depression, anxiety and mania at the symptom level in Latinos taking computerized adaptive tests in Spanish. We determined which symptoms that were good discriminators of high and low severity in a non-Latino population taking the tests in English were poor discriminators in a Spanish speaking Latino population. The findings provide relatively unbiased cross-cultural psychiatric comparisons.



Panel a. Depression



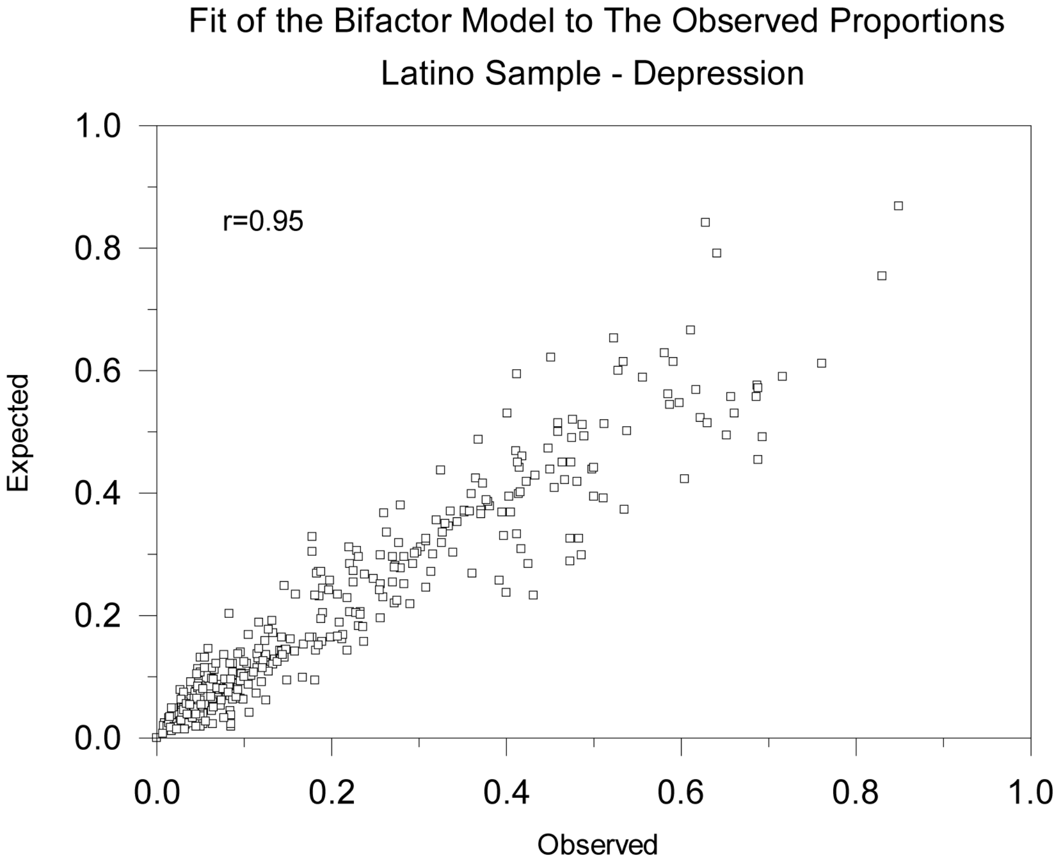
Panel b. Anxiety



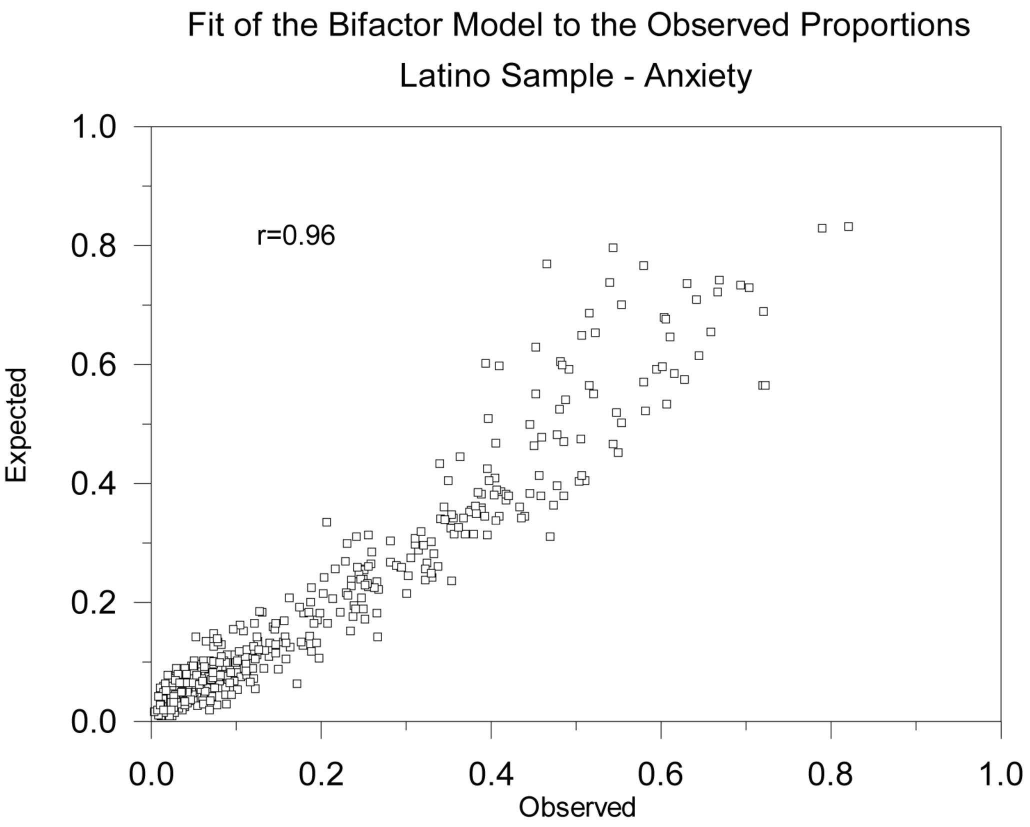
Panel c. Mania/Hypomania

Figure 1.

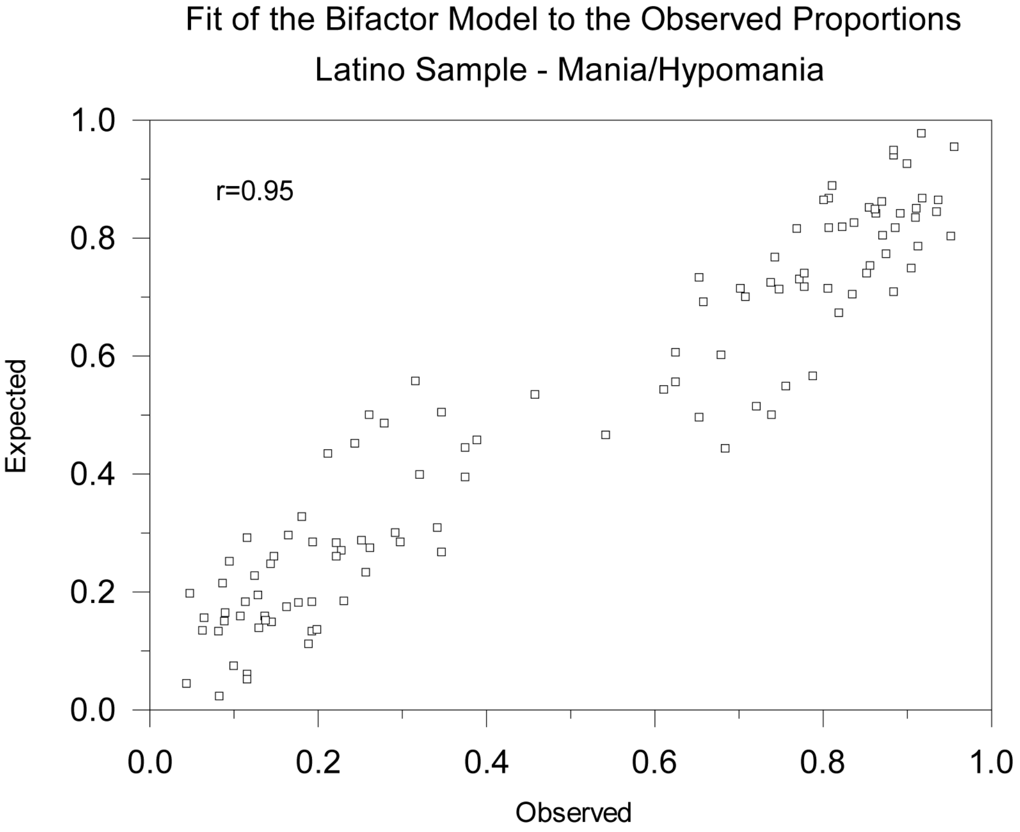
Bivariate relationship between the estimated primary severity scores based on the original non-Latino calibration and the new Latino calibration after eliminating poor discriminating items. X-axis refers to the primary severity score estimate based on the original non-Latino calibration. Y-axis refers to the primary severity score estimate based on the Latino calibration. The scale is the underlying N(0,1) distribution of the test scores produced by the bifactor model.



Panel a. Depression



Panel b. Anxiety



Panel c. Mania/Hypomania

Figure 2.

Model fit. Observed and expected marginal response category proportions based on the bifactor model for the Latino sample data. Panel (a) displays the 377 categories for the 77 depression items most of which had 5 categories. Panel (b) displays the 400 categories for the 84 anxiety items most of which had 5 categories. Panel (c) displays the 110 categories for the 55 mania/hypomania items most of which had 2 categories.

Table 1

Sociodemographic Breakdown of the Latino Sample

Sociodemographics	Latino %	Non-Latino %
Age		
18–34	39.4	29.5
35–49	37.2	31.5
50+	23.4	39.0
Gender		
Male	33.0	30.0
Female	67.0	70.0
Ethnicity Latino	100.0	1.0
Race		
White	25.0	45.0
Black	4.2	24.0
Hispanic/Latino/Caribbean	17.0	1.0
Mixed Race	43.3	2.0
Other	7.2	1.0
Missing	3.3	27.0
Education level		
Less than High School	40.5	5.0
HS Diploma, GED, Vocational School, or More	59.5	95.0

Table 2:

English and Spanish version of the DIF Depression and Mania Items where there were differences.

Depression Items	
English version	**Spanish version**
In the past 2 weeks, how much have you felt cheerful?	En las últimas dos semanas, ¿se ha sentido animado(a)?
In the past 2 weeks, I felt satisfied.	En las últimas dos semanas, ¿se sintió satisfecho(a)?
In the past 2 weeks, how easily did you get tired?	En las ultimas dos semanas, ¿con que facilidad se canso?
In the past 2 weeks, I had trouble keeping my mind on what I was doing.	En las últimas dos semanas, ¿cuán frecuentemente tuvo problemas para concentrarse en lo que estaba haciendo?
Mania Items	
English Version	**Spanish version**
In the past 2 weeks, did you ever engage in risk-taking behaviors, such as driving fast, promiscuous sex, hanging out in dangerous neighborhoods ?	En las últimas dos semanas, ¿alguna vez participó en conductas de alto riesgo, como conducir rápido, tener sexo promiscuo, o ir a barrios peligrosos?
In the past 2 weeks, have you had periods of at least 3 days in which you were preoccupied with yourself and your own problems, thoughts, and feelings?	En las últimas dos semanas, ¿ha tenido períodos de al menos 3 días en los que se preocupaba por si mismo(a) y por sus propios problemas, pensamientos y sentimientos?
In the past 2 weeks, have you had periods of at least 3 days in which you felt self-assured, charismatic or tended to assume a leadership role?	En las últimas dos semanas, ¿ha tenido períodos de al menos 3 días en los que se sentía seguro(a) de sí mismo, carismático(a) o tendía a asumir un papel de liderazgo?
In the past 2 weeks, have you had periods of at least 3 days in which you were less sexually active than is typical for you?	En las últimas dos semanas, ¿ha tenido períodos de al menos 3 días en los que se sentía seguro(a) de sí mismo, carismático(a) o tendía a asumir un papel de liderazgo?