Lessons learned from the clinical reappraisal study of the Composite International Diagnostic Interview with Latinos

MARGARITA ALEGRIA,[1,2] PATRICK E. SHROUT,[3] MARIA TORRES,[4] ROBERTO LEWIS-FERNÁNDEZ,[5] JAMIE M. ABELSON,[6] MERIS POWELL,[7] ALEJANDRO INTERIAN,[8] JULIA LIN,[1,2] MARA LADERMAN[1] & GLORISA CANINO[9]

1 Center for Multicultural Mental Health Research, Cambridge Health Alliance, Cambridge, MA, USA
2 Department of Psychiatry, Harvard Medical School, Boston, MA, USA
3 Department of Psychology, New York University, New York, NY, USA
4 The Heller School for Social Policy and Management, Brandeis University, Waltham, MA, USA
5 Department of Psychiatry, Columbia University and New York State Psychiatric Institute, New York, NY, USA
6 Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
7 Clinician in Private Practice and Senior Research Interviewer, New York, NY, USA
8 Department of Psychiatry, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ, USA
9 Behavioral Sciences Research Institute, University of Puerto Rico, San Juan, Puerto Rico

Abstract

Given recent adaptations of the World Health Organization's World Mental Health Composite International Diagnostic Interview (WMH-CIDI), new methodological studies are needed to evaluate the concordance of CIDI diagnoses with clinical diagnostic interviews. This paper summarizes lessons learned from a clinical reappraisal study done with US Latinos. We compare CIDI diagnoses with independent clinical diagnosis using the World Mental Health Structured Clinical Interview for DSM-IV (WMH-SCID 2000). Three sub-samples stratified by diagnostic status (CIDI positive, CIDI negative, or CIDI sub-threshold for a disorder) based on nine disorders were randomly selected for a telephone re-interview using the SCID. We calculated sensitivity, specificity, and weight-adjusted Cohen's kappa. Weighted 12 month prevalence estimates of the SCID are slightly higher than those of the CIDI for generalized anxiety disorder, alcohol abuse/dependence, and drug abuse/dependence. For Latinos, CIDI-SCID concordance at the aggregate disorder level is comparable, albeit lower, to other published reports. The CIDI does very well identifying negative cases and classifying disorders at the aggregate level. Good concordance was also found for major depressive episode and panic disorder. Yet, our data suggests that the CIDI presents problems for assessing post-traumatic stress disorder (PTSD) and generalized anxiety disorder (GAD). Recommendations on how to improve future versions of the CIDI for Latinos are offered. *Copyright © 2009 John Wiley & Sons, Ltd.*

Introduction

Structured diagnostic interviews are used in large-scale surveys as a way to eliminate clinician bias, improve standardization of diagnosis, and assess prevalence of psychiatric disorders in national studies (Komiti *et al.*, 2001). The World Health Organization (WHO) Composite International Diagnostic Interview Version 3.0 (CIDI 3.0; Kessler and Ustun, 2004) is a standardized diagnostic interview designed to assess current and lifetime mental disorders according to the definitions and criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; APA, 1994). The primary features of the CIDI are its reliance on respondents' self-report and standardization of administration by reliable non-clinicians after a relatively brief training period (Wittchen, 1994).

As of 2000, prior versions of the CIDI had been administered to more than 400 000 individuals in various studies (Andrews, 2000). Most recently, the CIDI was adapted for use in the Collaborative Psychiatric Epidemiology Surveys (CPES; Colpe *et al.*, 2004), a collection of three surveys that obtained psychiatric epidemiological information on mental disorders and service use in the US general population, with special emphasis on minority groups. It includes the National Comorbidity Survey Replication (NCS-R; Kessler and Merikangas, 2004), the National Latino and Asian American Study (NLAAS; Alegría *et al.*, 2004); and the National Survey of American Life (NSAL; Jackson *et al.*, 2004).

A clinical reappraisal phase was built into the design of all three CPES surveys and in several of World Mental Health (WMH) countries surveyed in order to evaluate the concordance of CIDI diagnoses with clinical diagnostic interviews. This paper summarizes the lessons learned from the clinical reappraisal study done with Latinos in the US and compares the results with those obtained from the WMH study (Haro *et al.*, 2006). We compare CIDI diagnoses with the independent clinical diagnoses obtained using the Structured Clinical Interview for DSM-IV (SCID; First *et al.*, 1998) as a way to interpret prevalence estimates generated for the Latino population and increase the clinical relevance of the CIDI in community surveys (Kessler *et al.*, 2004). We also conduct an individual-level examination across instruments to identify main reasons for discordance in diagnoses.

Background

Validation of the CIDI is a challenging task requiring comparison of CIDI results to psychiatric diagnoses similar to those a clinician would make under ideal circumstances (Kurdyak and Gnam, 2005). However, due to the highly unreliable nature of clinician assessments (Mellsop *et al.*, 1982; Spitzer and Fleiss, 1974), the use of clinician diagnoses as a validating gold standard has been questioned (Andrews and Peters, 1998). Clinical reappraisal studies of the CIDI have compared CIDI diagnostic results to clinician-administered semi-structured diagnostic interviews, such as the SCID, that allow the clinician some latitude in interpreting and coding responses (Kurdyak and Gnam, 2005). More recently, the WMH reappraisal studies (Haro *et al.*, 2006) examined whether the diagnoses in the CIDI are consistent with the SCID without assuming that one of the instruments is 'correct' (Kessler *et al.*, 2004). The NLAAS clinical reappraisal makes the same assumption.

Previous reappraisal studies have shown inconsistent findings with earlier versions of the CIDI based on DSM-III (third edition) and DSM III-R (revised third edition) criteria (Andrews, 2000; Brugha *et al.*, 2001; Janca *et al.*, 1992; Peters and Andrews, 1995). The results of these past clinical reappraisals varied according to the instrument and methods used to calibrate the CIDI, as well as the type of population surveyed (community versus clinic) and the time frame examined (last month versus lifetime). In studies in which the calibration method is not blind to the results of the CIDI interview, and lifetime rates are compared, the concordance between the CIDI and the external criterion is usually better. For example, in a clinical study comparing a clinical checklist with the CIDI after the clinicians were allowed to observe the CIDI interview (Wittchen, 1994), the overall lifetime kappa was high (0.77). However, in another study (WHO, 1995) using a clinical sample in which the CIDI was compared with the Schedules for Clinical Assessment of Neuropsychiatry (SCAN) with the clinician blind to CIDI probes, kappas ranged from poor (0.17) to moderate (0.61) (Andrews *et al.*, 1995). Concordance has usually been found to be lower when a population-based sample is used, when last-year rates are calculated rather than lifetime rates, and when the calibration method employed is blind to the results of the CIDI. In a large-scale, population-based study that compared last-month CIDI diagnoses with the SCAN, the results showed that with the exception of social phobia ($\kappa = 0.41$) and dysthymia ($\kappa = 0.48$), poor concordance was observed across most current psychiatric disorders (kappas ranged from -0.03 for GAD to 0.38 for any phobia; Brugha *et al.*, 2001). However, these past validation studies have been criticized for their limited examination of community samples and limited generalizability to more recent versions of the CIDI (Kurdyak and Gnam, 2005).

Clinical reappraisal data on the most recent version of the CIDI (the DSM IV-CIDI 3.0) is limited to two studies: the WHM study (Haro *et al.*, 2006) and the NCS-R (Kessler *et al.*, 2004), which essentially used the same methodology. Rather than follow the double-blind design of conventional clinical reappraisal studies, these studies unblinded the clinical interviewers to whether the respondents endorsed diagnostic stem questions in the CIDI, but not to the final CIDI diagnoses. They encouraged respondents to endorse diagnostic stem questions in the clinical reappraisal interviews by reminding respondents who endorsed CIDI stem questions of this fact (Haro *et al.*, 2006). The researchers argued that although this partial unblinding may introduce bias, this was likely not an issue due to the fact that the majority of community respondents who endorse CIDI stem questions do not go on to meet full CIDI criteria for the associated disorder (Kessler *et al.*, 2004). Therefore, results of both the WMH and NCS-R studies generally show better concordance between the CIDI and the SCID than other studies. For example, the NCS-R clinical reappraisal study showed a lifetime overall kappa of 0.53, with specific diagnoses ranging from a low of $\kappa = 0.35$ for social phobia to a high of $\kappa = 0.81$ for alcohol abuse (Kessler *et al.*, 2004). The WHM study indicated a generally good lifetime CIDI-SCID agreement (most kappas >0.40). Last-year agreement between the two instruments was generally good, but was presented only for modified aggregated diagnoses (i.e. anxiety, mood, and any disorder), which excluded the diagnoses of generalized anxiety disorder (GAD), dysthymia and post-traumatic stress disorder (PTSD).

The purpose of our study was to estimate the concordance of diagnoses for Latinos living in the US based on CIDI 3.0 (Kessler and Ustun, 2004) with diagnoses based on a follow-up clinical interview schedule, a slightly modified non-patient edition of the SCID (First *et al.*, 1998). However, this study did not unblind respondents or clinicians to any CIDI diagnostic information because of concern for unquantifiable bias that results from such information sharing. The results of this study are expected to add to the current literature on CIDI-SCID concordance by providing results from a diverse, multilingual community sample of Latinos living in the US.

Methods

Sample

The sample consists of Latino adults 18 years and older who completed an interview for the NLAAS. Latinos are persons of Latin-American or Spanish-speaking descent that self identified as Latino. Identification of psychiatric disorders was evaluated using the diagnostic interview of the WMH Survey Initiative version of the CIDI. The CIDI is a fully structured diagnostic instrument based on DSM-IV criteria that is administered by trained lay interviewers. Diagnoses were determined for 13 DSM-IV disorders but only nine were included for the clinical reappraisal: major depressive episode, dysthymia, agoraphobia without panic disorder, panic disorder, GAD, social phobia, PTSD, and alcohol abuse and/or dependence and drug abuse and/or dependence.

Three randomly selected sub-samples stratified by diagnostic status (CIDI positive, CIDI negative, or CIDI sub-threshold for a disorder), based on diagnosis for the above-referenced nine disorders, were randomly selected for a telephone re-interview by trained clinical interviewers using the SCID. Both English and Spanish-speaking respondents who met the selection criteria were eligible for inclusion in the study. The sample was drawn from different stages of the interview process (middle and late) to control for potential biases in the time of the year when the CIDI was conducted. Data collection was conducted from December 2002 through October 2003. During that time, 632 cases were randomly drawn from the primary NLAAS Latino sample based on diagnoses of the nine disorder categories. Of these 632 cases, 307 were selected for positive diagnosis, 124 were selected for sub-threshold diagnosis, and 201 were selected for negative diagnosis. The procedure required respondents selected for re-interview to be blind to their diagnostic status when their contact information was sent to the research group performing the interviews. Respondents were re-contacted and interviewed within 6–8 weeks of the CIDI interview. During this short period of time, a new request for an interview was made, followed by at least one additional telephone appointment for interviewing. If no contact was made within that timeframe, the interview was considered invalid. The data collection phase lasted 40 weeks, with a final sample of 195 respondents. However, less than 40% of respondents selected were able to be re-contacted ($n = 240$) within the allowable time period. Of those respondents who had operating phone numbers, were at home during times we attempted to contact them and re-contact happened within the 6–8 week time period, 81.3% agreed to a clinical re-interview.

Of these 195 cases, 48 respondents had at least one CIDI positive diagnosis among one of the nine assessed disorders, 59 were from the CIDI sub-threshold diagnosis group, and 88 were from the CIDI negative group. When we compared those whom we were able to re-contact from those we were not, we found immigrant Latinos and male

respondents to be less likely to be re-contacted than women and US born respondents.

Data collection

Trained clinicians administered the World Mental Health Structured Clinical Interview for DSM-IV (WMH-SCID 2000) via telephone to the reappraisal sample. This semi-structured clinical interview assesses for diagnosis in the context of specific DSM-IV Axis I disorders (First *et al.*, 1998). The World Mental Health Composite International Diagnostic Interview (WMH-CIDI) was translated into Spanish and other languages in combination with a major cross-national study of mental health by the WHO (World Mental Health Survey Consortium, 2004). For the NLAAS survey, the Spanish version of the core instruments also underwent further evaluation following the procedures described by Bravo *et al.* (1991). A full Spanish translation and adaptation of the SCID 2000 was completed using a back translation method and study-specific materials were developed for this study. Of the 195 CIDI interviews, 87 were conducted in English and 108 in Spanish.

Each interviewer received an initial 50+ hours of training on administering the SCID using BIOMETRIC tapes, group discussions, group practices, and reliability testing, followed by ongoing supervisory group interviewer meetings. Respondents were invited to conduct the reappraisal in their language of choice. Spanish-speaking respondents were interviewed by trained staff at the Department of Psychiatry, Robert Wood Johnson Medical School (RWJMS) while English-speaking respondents were interviewed by staff at the University of Michigan Institute for Social Research. These clinicians interviewed respondents with the SCID following the same supervisory structure described by Kessler *et al.* (2004). However, our procedure required that both clinicians and patients selected be blind to the respondent's diagnostic screening question status when their contact information and names were sent to the SCID interviewers. To control for order effects in the administration of the SCID by the clinician, respondents chosen for reappraisal were randomly placed in three ordered diagnostic patterns starting with anxiety, affective disorder, or substance use. The selection of the rotated SCID was decided randomly using the last number of the identification code of the respondent. Respondents who answered positively to the screening questions in the SCID were administered the corresponding diagnostic sections. Because of the semi-structured format of the SCID, the clinician could probe further if there was any suggestion that criteria were present beyond the answers obtained from the screening questions.

Initial contacts with respondents were made by non-clinicians to secure verbal informed consent for re-interview; once consent was received, a written consent form was sent to respondents and they were re-contacted by clinicians to conduct the interview. Payment of $50 was sent to respondents upon completion of the interview. All interviewers and supervisors were trained by supervisors who worked on the clinical reappraisal study of the NCS-R. Throughout the interviewing process, there were two levels of SCID interview quality control. First, a local interview supervisor checked the quality of the first 10 interviews for each interviewer and 25% of all interviews thereafter using audiotapes. Second, these previously reviewed interviews were re-reviewed by the offsite clinical interview supervisor who provided feedback to the local interview supervisor and, if required, assisted local interviewers in recalibrating their SCID ratings if 'drift' was detected in rater accuracy.

Data entry and analysis

Analysis focused on estimating the concordance between the CIDI and SCID for the selected disorders. The first step in this process was the creation of diagnostic algorithms for the SCID, following the procedures reported by Haro *et al.* (2006). Using the existing CIDI algorithms and the DSM-IV as a guide, two clinical psychiatrists expert in Latino diagnoses, in consultation with Jamie Abelson at the University of Michigan, developed the SCID algorithms for all nine disorders. Once the algorithms had been created and tested, CIDI disorder diagnoses were merged with SCID diagnoses. Twelve month disorder diagnoses were used for both the CIDI and the SCID.

Statistical analyses

We analyzed concordance between CIDI and SCID diagnoses after adjusting for the sampling design, which oversampled positive CIDI cases. We adjust for oversampling by reweighting the cases so that the overall prevalence rate matches that of the population. To assess concordance, we report sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) . These provide complementary information about the conditional probability of diagnosis using one method given a diagnosis using the other method. Neither method is a bona fide criterion, but we define sensitivity/specificity conditioning on the SCID and we define PPV/NPV conditioning on the CIDI (Fleiss *et al.*, 2003). Like Haro *et al.* (2006), we report an estimate of the area under the signal detection curve (AUC) using the average of

sensitivity and specificity (AUC1). We also compute the complementary statistic as the average of PPV and NPV labeled AUC2 (AUC2 is interpreted as the area under the curve if the CIDI were considered to be the criterion). We calculate confidence intervals on weight-adjusted Cohen's kappa using the method of Feder (2007).

To explore influences on concordance, we carried out variations of the analyses including: (a) loosening the timeframe (lifetime versus last year) when the disorder was present; (b) loosening the diagnostic criteria to allow for one criterion less than those stipulated in the DSM-IV to fulfill criteria for the disorder; and (c) requiring additional measures of impairment or dysfunction to establish illness severity. It is important to note that the CIDI algorithm does not operationalize all DSM-IV criteria, thus we only observed differences in criteria that were operationalized in both the CIDI and SCID instruments. For example, for Major Depressive Episode, DSM-IV criteria B and E were not operationalized. Criterion B states that the symptoms do not meet criteria for a Mixed Episode, and Criterion E states that the symptoms cannot be better accounted for by Bereavement ['after the loss of a loved one, the symptoms persist for longer than two months or are characterized by marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation' (APA, 1994)].

Using the CIDI and SCID algorithms, we completed a case-by-case analysis assessing reasons for discordance in every case where the respondent met criteria for a specific disorder in one instrument, but not in the other instrument (i.e. positive for dysthymia in the CIDI interview, but not in the SCID interview, or vice versa). Clinician interviewers were asked to document any problems that arose while administering the SCID.

Results

Table 1 presents summary results on WMH-CIDI and SCID concordance for disorders in the past 12 months using diagnostic categories reported by Haro *et al.* (2006, p. 176) for the WHO studies. To make comparisons of NLAAS to WHO results explicit, Table 1 reprints results from the Haro *et al.* (2006) article. For any anxiety disorder, the NLAAS AUC1 was 0.72 and $\kappa = 0.41$ [95% confidence interval (CI) = 0.18, 0.65], while for any depressive disorder, the AUC1 was 0.67 and $\kappa = 0.38$ (95% CI bounds = 0.17, 0.60). These compare to WHO AUC1 and kappa results of 0.88 and $\kappa = 0.42$ for anxiety disorders and 0.83 and $\kappa = 0.56$ for depressive disorders. Any disorder in NLAAS had an AUC1 of 0.71 and $\kappa = 0.39$

Table 1 Consistency of twelve-month DSM-IV CIDI and SCID diagnoses in the ESEMeD (*n* = 143) and NLAAS (*n* = 195) clinical reappraisal samples

	AUC1[a]	κ	SE	OR	95% CI	SN	SE	SP	SE	PPV	SE	NPV	SE
ESEMeD any anxiety disorder	0.88	0.42	0.1	66.7	18.2–244.2	83.7	8.3	92.9	1.7	31.3	7.8	99.3	0.4
ESEMeD any mood disorder	0.83	0.56	0.2	76.7	21.4–274.9	69.1	11.8	97.2	0.9	49.6	10.5	98.7	0.6
ESEMeD any disorder	0.84	0.49	0.1	34.1	6.6–176.8	77.9	13.2	90.6	2.7	41.5	8.5	98.0	1.5
NLAAS any depressive disorder[b]	0.67	0.38	0.1	14.4	4.0–51.1	38.9	11.0	95.8	1.9	49.1	13.3	93.7	2.2
NLAAS any anxiety disorder[c]	0.72	0.41	0.1	22.9	6.3–83.0	48.8	13.7	96.0	1.4	41.4	12.5	97.0	1.2
NLAAS any disorder[d]	0.69	0.39	0.1	10.0	3.5–28.9	44.9	10.3	92.5	2.4	50.3	9.9	90.8	3.1

[a] AUC1 = mean (sensitivity, specificity).
[b] Exclude dysthymia.
[c] Exclude PTSD and GAD.
[d] Exclude dysthymia, PTSD and GAD.
Note: ESEMeD, European Study of the Epidemiology of Mental Disorders; SE, standard error; OR, odds ratio; SN, sensitivity; SP, specificity.

(95% CI = 0.20, 0.58), compared to WHO AUC1 and kappa results of 0.84 and 0.49, respectively. We conclude that the overall concordance of the NLAAS assessments was similar, albeit weaker, to the concordance between CIDI and SCID assessments in the WHO studies. The confidence bounds of the NLAAS results include the values reported by WHO investigators.

Detailed comparison of individual 12 month diagnoses reveals interesting patterns of discrepancy and convergence. Table 2 shows results for nine specific diagnoses. The AUC1 values range from 0.77 to 0.49, with a median of 0.61, and the kappa values range from 0.63 to −0.01, with a median of 0.18. The AUC2 values vary from 0.98 to 0.48, with a median of 0.70. For only two of the nine diagnoses does the level of agreement exceed what could be expected by chance. These are for major depressive episode (AUC = 0.67, κ = 0.38) and panic disorder (AUC = 0.77, κ = 0.63). Two of the remaining seven diagnoses (alcohol abuse/dependence, drug abuse/dependence) were assigned to fewer than five persons out of 195 by the CIDI, therefore the comparisons are inconclusive. In addition, the confidence bounds on the kappa statistic for social phobia (−0.04, 0.61), dysthymia (−0.14, 0.42), and agoraphobia (−0.03, 0.59) include both chance values and acceptable values, resulting in inconclusive comparisons. The final two diagnoses (PTSD and GAD) have sufficient precision to conclude that the data are inconsistent with kappa values of 0.50 or higher, indicating systematic discrepancies between the two assessment methods. For example, of 12 cases identified by one of the instruments as fulfilling criteria for PTSD, five were diagnosed by the CIDI and a different group of seven respondents were identified by the SCID. Likewise for GAD; of the 15 cases classified by one of the instrument as having the disorder, six were classified by the CIDI and a different group of seven respondents were identified by the SCID.

Because WMH-CIDI and SCID establish the time frame of the disorder in such different ways (e.g. SCID focuses on assessing last-year and then lifetime prevalence, while CIDI examines lifetime prevalence first and then establishes if disorder was present in the last year), it is possible that concordance would be different if a lifetime, rather than a last-year, time frame was used to compare WMH-CIDI and SCID results. We found (data not shown) that only for substance use disorders did the AUC1 considerably improve when using the lifetime time frame, becoming 0.82 for lifetime concordance rather than 0.56. There was also an increase in kappa, with kappa now reaching 0.51 (0.23, 0.80) for lifetime concordance as compared to κ = 0.18 (−0.05, 0.41) for last-year concordance.

We also conducted additional analyses allowing the diagnostic criteria to be met with one criterion less than those stipulated in the DSM-IV for both SCID and CIDI (data not shown). Kappa for any depressive episode improved from 0.38 to 0.43, and from 0.14 to 0.28 for dysthymia. Concordance for some anxiety disorders improved (GAD and PTSD) while others worsened (e.g. panic; data not shown). We also tested whether requiring dysfunction/impairment would improve concordance between CIDI and SCID. We found that limiting comparisons to diagnoses associated with a Global Assessment of Functioning (GAF) score of 60 or less in the SCID and one or more disability days in the CIDI improved concordance for social phobia, agoraphobia, alcohol abuse/dependence and drug abuse/dependence. Kappas increased from 0.28 to 0.61 for social phobia; from 0.18 to 0.64 for agoraphobia; from 0.16 to 0.26 for alcohol abuse/dependence; and from 0.32 to 0.39 for drug abuse/dependence).

We examined the reasons for discordance between the two instruments. Of 195 cases assessed, 110 (56%) evidenced discordance due to one of five reasons (Table 3). For example, 20 cases (18%) were discordant because one instrument coded all necessary criteria as present whereas the other instrument missed the diagnosis by only one criterion (category 1 on Table 3); this occurred primarily in the diagnoses of major depressive episode, dysthymia, and GAD. Fourteen other cases (13%) were discordant on two or more criteria, as seen for social phobia, GAD, major depressive episode or dysthymia (category 2). Twenty-five cases (23%) were discordant because in one instrument but not the other respondents skipped out of the battery (category 3) by answering 'no' to one of the probes required to continue into the next section. This was much less common in the SCID than the CIDI, where it happened in 23 of 25 cases. Another 14 cases (13%) fulfilled criteria for lifetime but not last-year disorder in the CIDI (category 4), but did fulfill last-year criteria in the SCID. An additional 37 cases (34%) were discordant because they did not endorse the screener probes in one instrument (25 in the SCID and 12 in the CIDI) but did endorse them in the other (category 5).

Upon reviewing the problems reported during SCID administration, interviewers often cited respondents' difficulties understanding questions and responding within a specified time frame. For major depressive episode, part 1 of Criterion A ('symptoms of depressed mood or loss of interest must happen within the same two-week period') proved to be problematic, accounting for discordance in 11 of 30 discordant cases. For dysthymia, 6 of 14 discordant cases evidenced problems with Criterion C ('two

Table 2 Consistency of twelve-month DSM-IV diagnoses of the CIDI and SCID for specific disorders assessed in the NLAAS clinical reappraisal (n = 195)

	Weighted CIDI Prevalence			Weighted SCID Prevalence			AUC1[a]	AUC2[a]	OR	95% CI
	n	Percentage	SE	n	Percentage	SE				
Major Depression	26	7.6%	2.1	32	9.5%	2.4	0.67	0.71	14.4	4.0–51.1
Dysthymia	7	1.6%	0.7	9	3.2%	1.3	0.55	0.61	10.7	1.0–111.1
Panic Disorder	5	1.4%	0.7	8	2.1%	1.0	0.77	0.89	375.2	24.5–5739.6
Social Phobia	15	3.9%	1.4	9	2.8%	1.2	0.67	0.62	19.0	2.9–125.0
Agoraphobia	5	1.0%	0.6	8	1.5%	0.7	0.62	0.68	48.3	4.9–477.7
GAD	6	1.2%	0.6	9	3.1%	1.3	0.49	0.48	0	0
PTSD	4	1.4%	0.6	5	1.1%	0.6	0.49	0.49	0	0
Alcohol Abuse/Dependence	2	0.4%	0.4	8	4.6%	2.2	0.55	0.98	–	–
Drug Abuse/Dependence	3	1.1%	0.6	6	3.0%	1.4	0.61	0.81	74.4	5.2–1061.4
Any Disorder[b]	44	14.0%	2.8	43	19.0%	3.6	0.70	0.75	12.0	4.4–32.9

	κ	95% CI	SE	SN	SE	SP	SE	PPV	SE	NPV	SE
Major Depression	0.38	0.17–0.60	0.1	38.9	11.0	95.8	1.9	49.1	13.3	93.7	2.2
Dysthymia	0.14	−0.14–0.42	0.1	12.0	11.6	98.7	0.6	23.8	20.0	97.2	1.3
Panic Disorder	0.63	0.24–1.03	0.2	53.7	23.8	100.0	0.3	78.5	19.6	99.0	0.7
Social Phobia	0.28	−0.04–0.61	0.2	36.4	20.1	97.1	1.2	26.6	15.7	98.1	1.0
Agoraphobia	0.28	−0.03 to −0.59	0.2	24.3	14.9	99.3	0.6	36.1	23.8	98.8	0.6
GAD	−0.02	−0.03 to −0.01	0	0	0	98.8	0.6	0	0	96.8	1.3
PTSD	−0.01	−0.02 to −0.01	0	0	0	98.6	0.6	0	0	98.9	0.6
Alcohol Abuse/Dependence	0.16	−0.11–0.43	0.1	9.2	8.4	100.0	0	100.0	0	95.8	2.1
Drug Abuse/Dependence	0.32	−0.06–0.70	0.2	22.5	15.8	99.6	0.4	64.5	28.1	97.6	1.3
Any Disorder	0.44	0.26–0.62	0.1	45.8	9.1	93.4	2.2	62.1	9.4	88.0	3.4

[a] AUC1 = mean (sensitivity, specificity); AUC2 = mean (PPV, NPV).
[b] Exclude PTSD and GAD.

Table 3 Detailed discordance report by categories/disorders ($n = 110$ discordant cases)

Discordance Categories	Major depression CIDI− SCID+	Major depression SCID− CIDI+	Dysthymia CIDI SCID+	Dysthymia SCID− CIDI+	GAD CIDI− SCID+	GAD SCID− CIDI+	Panic disorder CIDI− SCID+	Panic disorder SCID− CIDI+	Agoraphobia CIDI− SCID+	Agoraphobia SCID− CIDI+	Social phobia CIDI− SCID+	Social phobia SCID− CIDI+	PTSD CIDI− SCID+	PTSD SCID− CIDI+	Alcohol abuse/ dependence CIDI− SCID+	Alcohol abuse/ dependence SCID− CIDI+	Drug abuse/ dependence CIDI− SCID+	Drug abuse/ dependence SCID− CIDI+	Total by categories
1. Missed by one criteria (have respondent information for all diagnostic criteria)	1	4	3	1	3	2						1	1	1	1		2	1	20
2. Missed by more than one criteria (have respondent information for all diagnostic criteria)		2	2		1	3			1			3		1	1				14
3. Entered battery and then skipped out/missed by at least one criteria (CIDI or SCID stopped inquiry regarding any remaining criteria when this criteria was not met)	8	2	2		3		2						4		2		2		25
4. Lifetime, but not 12 months in CIDI	8		1		2		1								1		1		14
5. Did not screen into diagnostic battery	1	4		5		1	1	1	4	2	4	6		5	2				37
Total negative cases by instrument	18	12	8	6	9	6	4	1	5	2	4	10	5	7	7	0	5	1	
Total discordant cases by disorder	30		14		15		5		7		14		12		7		6		110

years in a row without a break in depressive symptoms lasting longer than two months or more'). For GAD, the criterion requiring that respondents identify their difficulty controlling worry was responsible for discordance in 5 of 15 cases. For alcohol and drug use disorders, Criterion A, dealing with whether the behavior is maladaptive and leads to significant impairment, was responsible for three of seven discordant cases for alcohol abuse/dependence and four of six discordant cases for drug abuse/dependence.

Discussion

The results of this study should be considered in light of several methodological limitations. Our findings are constrained by small sample sizes for five of the nine assessed disorders. Fewer than five cases received CIDI diagnoses of either alcohol abuse/dependence or drug abuse/dependence and confidence bounds on the kappas for social phobia, dysthymia, and agoraphobia suggested inconclusive results. In addition, not all criteria were assessed by the CIDI, so results may vary if this additional information (e.g. collecting information on bereavement) were collected during assessment. Furthermore, some diagnoses that are part of DSM-IV (e.g. mania, schizoaffective disorders, somatisation disorders) were not assessed in the NLAAS or the clinical reappraisal datasets, and therefore were excluded from these analyses. We did not include somatisation because of lack of consensus on the criteria to assess the disorder.

Notwithstanding, given the recent adaptation of the CIDI for use in the CPES, our study evaluates the concordance of CIDI diagnoses with clinical interviews obtained with the SCID. For Latinos, CIDI-SCID concordance at the aggregate level of disorder is comparable to, although lower than, the other published reports on the CIDI 3.0, as described by Haro *et al.* (2006). The lower kappas would be expected due to the fact that the NLAAS clinical reappraisal followed a blinded design and Haro *et al.* (2006) reported reappraisal studies did not. As can be seen from the results in Table 3, 13% of the discordance (25/195 discordant cases) was due to respondents denying during the later SCID interview that they had ever endorsed the CIDI screener probes.

The classificatory accuracy – or caseness established – ranged from 78% for any disorder to 95% for substance use disorders for our Latino sample. For Latinos, the CIDI does very well identifying negative cases and classifying disorders at the aggregate level. Good concordance was also found for major depressive episode and panic disorder. This may be due to the fact that panic is a paroxys-

mic, discrete event with high valence and that major depressive episode is a familiar condition with symptoms that are easy to comprehend and describe, facilitated by public awareness campaigns about its symptoms (Correll and Linden, 2005).

However, our data suggest that the CIDI 3.0 presents problems for assessing PTSD and GAD, and also needs additional testing regarding social phobia, dysthymia and agoraphobia. The lack of overlap between the CIDI and the SCID assessments of PTSD was dramatic, probably due to the different ways of evaluating the presence of the disorder across instruments. The CIDI assesses a long list of specific traumatic events, whereas the SCID asks a general question about trauma exposure. Other potential explanations for the observed lack of concordance relate to characteristics of the illness itself, which includes avoidance as one of its symptoms. PTSD patients might want to try to avoid recounting their symptoms, particularly soon after being asked to remember these events. An even more likely explanation is that exposure (both in terms of severity and chronicity of the exposure) and PTSD characteristics may differ significantly by race and ethnicity (Antai-Otong, 2002; Elsass, 2001; Hernandez, 2002; Hernandez *et al.*, 2007), with clinical assessment being particularly problematic for clients from non-Western backgrounds. Because the evaluation of trauma in the SCID, as opposed to the CIDI, is based on the clinician's own interpretation of the meaning of a particular trauma given the social norms of the patient's culture, it might be particularly challenging for clinicians to assess PTSD in the absence of cultural anchors (Alarcon, 1995). So if the patient says that witnessing domestic violence is typically expected in his/her home country, the clinician might judge that the event was not traumatic for the respondent, while the CIDI structure does not allow the interviewer to assess the impact of the trauma based on how normative is the experience in the respondent's context.

For GAD, our low concordance results are consistent with the literature on previous versions of the CIDI that suggest low levels of sensitivity for detecting GAD (Komiti *et al.*, 2001). In examining the procedural validity of CIDI diagnoses of GAD, Wittchen *et al.* (1995) found that GAD diagnoses obtained with the UM-CIDI (a modified version of the CIDI used in the NCS) also showed low levels of agreement with SCID diagnoses ($\kappa = 0.35$). Future work is needed to address whether Latino respondents have particular difficulty understanding or endorsing the criterion requiring respondents to identify difficulty controlling their worry as a condition of a GAD diagnosis.

Our results suggest that, for Latinos, loosening the diagnostic criteria for each category by one item might improve CIDI–SCID concordance for depressive disorders and some anxiety disorders (GAD and PTSD). This finding requires further inquiry. It is possible that certain diagnostic criteria do not apply as well to Latinos because they represent a category fallacy (Kleinman, 1987), whereby concepts used in one culture do not map easily onto another culture. For instance, questions about the duration requirement for major depressive episode (e.g. depressed mood or loss of interest must happen within the same two week period with other symptoms) might not correspond to Latino time concepts about depressive illness leading to inconsistent answers depending on how the questions are asked. Assessing conceptual equivalency for monolingual Spanish speakers, particularly immigrants, may help clarify whether lack of endorsements of probes happens more readily when Latinos do not ascribe to mental health concepts in the same way as set out in the DSM-IV. Given the insufficient sample size in our Clinical Reappraisal study, we were unable to evaluate the conceptual equivalence for monolingual Spanish speakers as compared to the monolingual English speakers and respondents that spoke both languages. A more extensive analysis, such as item response theory analysis, would be required to tease out this potential measurement bias from differences in endorsement rates due to severity of depressive symptoms.

An alternative explanation involves difficulties understanding certain concepts embedded in the questions, possibly due to educational barriers or cultural differences. These include the illness episode concept or the evaluative element embedded in certain questions requiring respondents to conduct a comparative assessment, such as deciding whether the behavior is maladaptive and leads to significant impairment. Because the SCID permits the clinician to return to or revise sections if interviewees disclose information relevant to a previous diagnostic module, further probes could have facilitated revising respondents' answers. Such is not the case in the CIDI, where 23 out of 25 respondents were dropped from the diagnostic battery because they denied one of the probes required to continue into the next part of the diagnostic section.

Loosening the time dimension also appears important particularly for substance use disorders, where concepts of time might be different in Latino culture (Canino *et al.*, 2004). Simplifying the criteria so as to decrease the salience of time might improve the clinical concordance of the CIDI for Latino respondents. The finding that SCID prevalence rates of substance use disorders are higher than those obtained by the CIDI, might also be due to clinicians' greater ability to elicit stigmatized and illegal behaviors. In addition, the SCID and CIDI assess substance use disorders differently, possibly contributing to the striking discrepancy in prevalence rates obtained by the two instruments. In the CIDI, respondents must meet abuse to be evaluated for dependence, whereas in the SCID, abuse and dependence are evaluated independently (see Grant *et al.*, 1996).

Requiring more stringent evaluations of dysfunction/impairment would likely improve the concordance between the CIDI and the SCID for social phobia, alcohol abuse/dependence, and drug abuse/dependence. Our data suggests that 'gray' cases, those respondents who were positive to one of the instruments and not the other, would particularly benefit from additional information on dysfunction to help establish their caseness status. The open-endedness of the SCID evaluation probably allowed clinicians to return to previously completed modules during the same evaluation and include the newly obtained material, which could not be done on the CIDI evaluation.

These reasons for discordance described earlier, combined with the fact that the aggregate disorders have a higher concordance than the individual disorders, suggest that discordance is due to methodological differences, particularly in how these instruments codify 'gray' cases. Certain diagnostic criteria seem to pose difficulty for Latino respondents, and should be better operationalized to avoid misclassifications depending on how the questions are asked. Although in both assessments Latino respondents appear to recognize the experience being assessed in either instrument, they either experience the disorder with minor variations (e.g. differing by one criterion) or have difficulty being exact about the time frame in one of the assessments. This variation might be better captured by a dimensional approach to diagnosis currently proposed for the DSM-V.

Our results suggest several methodological improvements to diagnostic assessments for Latino respondents. These include expanding screening questions, opening up the time frames for assessment, gathering more information within each diagnostic section prior to skipping out respondents, clarifying the threshold for severity or dysfunction, and loosening the criteria for certain disorders that appear not to map so readily onto illness expressions typical among Latinos. Populations differ in terms of illness expressions, sense of time, health literacy, and rates of formal education. This variation could affect how they answer questions on structured and semi-structured instruments. A goal of epidemiology is to assess variations in psychopathology across population subgroups,

which requires that enough information be obtained on each disorder to evaluate its variation. One of the major problems faced by the field stems from limitations of current instruments (e.g. skip patterns, limited symptom inclusion) which do not permit an easy comparison of alternative illness expressions across groups. The proposed improvements should be tested in future studies to evaluate their impact on concordance. Improved concordance between diagnostic assessments will aid in the interpretation of DSM-IV prevalence estimates generated by the CIDI and increase the clinical relevance of the CIDI for community epidemiological surveys (Kessler *et al.*, 2004).

Acknowledgments

Declaration of interest statement

The authors have no competing interests.

References

Alarcon R.D. (1995) Culture and psychiatric diagnosis: Impact on DSM-IV and ICD-10. *The Psychiatric Clinics of North America*, **18**, 449–465.

Alegría M., Takeuchi D., Canino G., Duan N., Shrout P., Meng X.-L. *et al.* (2004) Considering context, place and culture: The National Latino and Asian American Study. *International Journal of Methods in Psychiatric Research*, **13**(4), 208–220.

American Psychiatric Association (APA). (1994) *Diagnostic and Statistical Manual of Mental Disorders – 4th Edition*, American Psychiatric Association.

Andrews G. (2000) Case ascertainment: The Composite International Diagnostic Interview. *Australian and New Zealand Journal of Psychiatry*, **34**, S161–S163.

Andrews G., Peters L. (1998) The psychometric properties of the Composite International Diagnostic Interview. *Social Psychiatry & Psychiatric Epidemiology*, **33**, 80–88.

Andrews G., Peters L., Guzman A., Bird K. (1995) A comparison of two structured diagnostic interviews: CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry*, **29**, 124–132.

Antai-Otong D. (2002) Culture and traumatic events. *Journal of the American Psychiatric Nurses Association*, **8**(6), 203–208.

Bravo M., Canino G., Rubio-Stipec M., Woodbury M. (1991) A cross-cultural adaptation of a diagnostic instrument: The DIS adaptation in Puerto Rico. *Culture, Medicine and Psychiatry*, **15**, 1–18.

Brugha T., Jenkins R., Taub N., Meltzer H., Bebbington P. (2001) A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychological Medicine*, **31**, 1001–1013.

Canino G., Shrout P.E., Rubio-Stipec M., Bird H.R., Bravo M., Ramirez R. *et al.* (2004) The DSM-IV rates of child and adolescent disorders in Puerto Rico: Prevalence, correlates, service use, and the effects of impairment. *Archives of General Psychiatry*, **61**(1), 85–93.

Colpe L., Merikangas K., Cuthbert B., Bourdon K. (2004) Guest editorial. *International Journal of Methods in Psychiatric Research*, **13**(4), 193–194.

Correll C., Linden M. (2005) Psychotropic drug presentation in medical and lay press journals. *Pharmacopsychiatry*, **38**(4), 161–165.

Elsass P. (2001) Individual and collective traumatic memories: A qualitative study of post-traumatic stress disorder symptoms in two Latin American localities. *Transcult Psychiatry*, **38**(3), 306–316.

Feder M. (2007) *Variance Estimation of the Survey-weighted Kappa Measure of Agreement*, JSM Proceedings.

First M.B., Spitzer R.L., Gibbon M., Williams J. (1998) *Structured Clinical Interview for DSM-IV Axis I Disorders*, Patient Edition (SCID-I/P, Version 2.0, 9/98 revision), Biometrics Research Department, New York State Research Institute.

Fleiss J.L., Levin B., Paik M.C. (2003) *Statistical Methods for Rates and Proportions, 4th Edition*, Hoboken, NJ: John Wiley & Sons.

Grant B. (1996) Prevalence and correlates of drug use and DSM-IV drug dependence in the United States: Results of the National Longitudinal Alcohol Epidemiologic Survey. *Journal of Substance Abuse*, **8**(2), 195–210.

Haro J., Arbabzadeh-Bouchez S., Brugha T., Girolamo G., Guyer M., Jin R. *et al.* (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys. *International Journal of Methods in Psychiatric Research*, **15**(4), 167–180.

Hernandez P. (2002) Trauma in war and political persecution: Expanding the concept. *American Journal of Orthopsychiatry*, **72**(1), 16–25.

Hernandez P., Gangsei D., Engstrom D. (2007) Vicarious resilience: A new concept in work with those who survive trauma. *Family Process*, **46**(2), 229–241.

Jackson J., Torres M., Caldwell C., Neighbors H., Nesse R., Taylor R.J. *et al.* (2004) The National Survey of American Life: A study of racial, ethnic and cultural influences on mental disorders and mental health. *International Journal of Methods in Psychiatric Research*, **13**(4), 196–207.

Janca A., Robins L., Bucholz K., Early T., Shayka J. (1992) Comparison of Composite International Diagnostic Interview and clinical DSM-III-R criteria checklist diagnoses. *Acta Psychiatrica Scandinavica*, **85**, 440–443.

Kessler R., Merikangas K. (2004) The National Comorbidity Survey Replication (NCS-R): Background and aims. *International Journal of Methods in Psychiatric Research*, **13**(2), 60–68.

Kessler R., Ustun T. (2004) The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, **13**(2), 93–121.

Kessler R., Abelson J., Demler O., Escobar J., Gibbon M., Guyer M. *et al.* (2004) Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Inverview (CIDI). *International Journal of Methods in Psychiatric Research*, **13**(2), 122–139.

Kleinman A. (1987) Anthropology and psychiatry: The role of culture in cross-cultural research on illness. *British Journal of Psychiatry*, **151**, 447–454.

Komiti A.A., Jackson H.J., Judd F.K., Cockram A.M., Kyrios M., Yeatman R. *et al.* (2001) A comparison of the Composite International Diagnostic Interview (CIDI-Auto) with clinical assessment in diagnosing mood and anxiety disorders. *Australian and New Zealand Journal of Psychiatry*, **35**, 224–230.

Kurdyak A., Gnam W. (2005) Small signal, big noise: Performance of the CIDI depression module. *Canadian Jounral of Psychiatry*, **50**(13), 851–856.

Mellsop G., Varghese F., Joshua S., Hicks A. (1982) The reliability of axis II of DSM-III. *American Journal of Psychiatry*, **139**, 1360–1361

Peters L., Andrews G. (1995) Procedural calidity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychological Medicine*, **25**(6), 1269–1280.

Spitzer R., Fleiss J. (1974) A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, **125**, 341–347.

Wittchen H. (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. *Psychiatry Research*, **28**, 57–84.

Wittchen H., Kessler R., Zhao S., Abelson J. (1995) Reliability and clinical validity of UM-CIDI DSM-III-R generalized anxiety disorder. *Journal of Psychiatric Research*, **29**(2), 95–110.

World Health Organization (WHO). (1995) *Composite International Diagnostic Interview (CIDI), version 2.1*, WHO.

World Mental Health Survey Consortium. (2004) Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Journal of the American Medical Association*, **291**(21), 2581–2590.