Psychological Assessment
2002, Vol. 14, No. 1, 50–59

1040-3590/02/$5.00 DOI: 10.1037//1040-3590.14.1.50

Differential Item Functioning in a Spanish Translation of the PTSD Checklist: Detection and Evaluation of Impact

Maria Orlando and Grant N. Marshall
RAND

This study demonstrated the application of an innovative item response theory (IRT) based approach to evaluating measurement equivalence, comparing a newly developed Spanish version of the Posttraumatic Stress Disorder Checklist–Civilian Version (PCL–C) with the established English version. Basic principles and practical issues faced in the application of IRT methods for instrument evaluation are discussed. Data were derived from a study of the mental health consequences of community violence in both Spanish speakers (n = 102) and English speakers (n = 284). Results of differential item functioning (DIF) analyses revealed that the 2 versions were not fully equivalent on an item-by-item basis in that 6 of the 17 items displayed uniform DIF. No bias was observed, however, at the level of the composite PCL–C scale score, indicating that the 2 language versions can be combined for scale-level analyses.

Latinos constitute a large and rapidly growing segment of the U.S. population, and recent migration has resulted in sizable percentages of Latin American immigrants and refugees with limited proficiency in English (Garcia-Preto, 1996). This trend has increased the use of Spanish-language versions of scales and entire surveys in large-scale survey research studies (e.g., Weidmer, Brown, & Garcia, 1999; Wells, 1999). Whereas extensive attention has been devoted to the translation process to ensure maximum validity before fielding the instruments (e.g., Brislin, 1970; Geisinger, 1994; Greenfield, 1997; Guthery & Lowe, 1972; Werner & Campbell, 1970), it is also important to examine the validity of the translated scale empirically (Budgell, Raju, & Quartetti, 1995). Whether the translation is the main focus of the study or simply a necessary component, the performance of the translated items and scales should be empirically examined before responses from different-language versions are assumed to have equivalent properties (Bjorner, Kreiner, Ware, Damsgaard, & Bech, 1998; Budgell et al., 1995).

As part of a prospective study of the mental health consequences of exposure to community violence, we developed a Spanish (Marshall & Jaycox, 2000) version of the Posttraumatic Stress Disorder Checklist—Civilian Version (PCL–C; Weathers, Litz, Herman, Huska, & Keane, 1993) for use with study participants who are not facile with the English language. Posttraumatic stress disorder (PTSD) is an anxiety disorder characterized by prolonged symptoms (30 days or more) of reexperiencing, avoidance, numbing, and hyperarousal after exposure to a traumatic event (American Psychiatric Association, 1994). Because participants in the current study were assessed within days of trauma exposure, we focus on PTSD-like symptoms rather than a PTSD diagnosis per se. To ensure the accuracy of cross-group comparisons and the appropriateness of collapsing responses across English- and Spanish-speaking participants for certain analytic purposes, we sought first to empirically establish that the English and Spanish versions of the PCL–C were measuring the same underlying construct to the same degree in our study sample.

Maria Orlando and Grant N. Marshall, RAND, Santa Monica, California.

This work was supported by National Institute of Mental Health Grant 1R01MH56122 and a grant from the William T. Grant Foundation.

Correspondence concerning this article should be addressed to Maria Orlando, RAND, 1700 Main Street, P.O. Box 2138, Santa Monica, California 90407-2138. E-mail: orlando@rand.org

Differential Item Functioning

Whereas classic correlation and reliability analyses can provide some information about scale equivalence (e.g., Negy & Snyder, 2000; Norris & Perilla, 1996), an alternative and highly instructive conceptualization of the nonequivalence of tests focuses on the presence of statistical item bias, or differential item functioning (DIF). Although DIF traditionally refers to an educational test item examined across gender or ethnic groups, the concept easily generalizes to include noneducational test items, such as items from a psychological scale, examined across nontraditional groupings, such as language of administration. In this context, an item is said to exhibit DIF if two respondents who are administered different-language versions of a scale and have equal levels of the psychological trait being measured do not have the same probability of endorsing each response category of that item. If, after controlling for overall group differences, the probabilities of endorsement are not equal for the two groups, the DIF is said to be uniform. Nonuniform DIF occurs when an item is more salient for one group than the other, so that the relationship between the item and the construct being measured is not equal for the two groups.

The practical result at the item level is that scores on an item exhibiting DIF are not equivalent across the groups being studied, leading to potentially misleading group differences and inaccurate bivariate associations involving the DIF item (Holland & Wainer, 1993). At the level of the scale, the impact of the presence of DIF items within a scale of items can vary depending on the degree of DIF, the number of items in the scale exhibiting DIF, and the proposed uses of the scale. Thus, impact must be evaluated on a situation-specific basis.

Classical Versus Item Response Theory Approaches to DIF Detection

A number of approaches to detecting DIF have been developed using methods from both classical test theory and item response theory (IRT; Holland & Wainer, 1993). Although classical test theory methods require few assumptions and are relatively easy to implement, results of these applications are sample specific and as such are not sufficient for ensuring measurement invariance (Budgell et al., 1995; Hulin, Drasgow, & Parsons, 1983). When the assumptions for an IRT application are met, this approach offers several advantages. Most notably, results from an IRT application generalize beyond the sample being studied to the population it represents.

The IRT approach also offers advantages with respect to interpretation and evaluation of DIF. Graphical representations of DIF and differential test functioning (DTF; Raju, van der Linden, & Fleer, 1995) can be created on the basis of results from an IRT analysis. These plots are valuable diagnostic tools for evaluating the potential impact of DIF both at the item level and at the level of the entire scale (DTF). Finally, in the IRT approach, once DIF is detected, scores can be generated that account for DIF. These scores can be compared with raw scores in sensitivity analyses to evaluate the practical impact of DIF on the application of interest.

DIF Detection Within IRT

Within the IRT framework, there are essentially two approaches to DIF analysis that vary both in the method of linking the two groups being studied— one of the more challenging aspects of DIF analysis—and in the identification of significant DIF. In the first general approach, item parameters obtained from separate calibrations of the two groups are equated using the total score as a basis for linking the groups (Stocking & Lord, 1983), and DIF is identified for each item using one or more DIF indices (Raju, 1990; Raju et al., 1995). Examples of this type of application can be found in Budgell et al. (1995), Collins, Raju, and Edwards (2000), Donovan, Drasgow, and Probst (2000), and Waller, Thompson, and Wenk (2000).

The second approach, illustrated in this article, calibrates the items for both groups simultaneously using IRT's "built-in" linking mechanism (Embretson, 1996) to link the two groups on the basis of a subset of items, referred to as anchor items, that are judged a priori to be unbiased. There are a number of ways to do this. In large-scale educational testing, anchor items are often selected out of a pool of established unbiased items. When prior information about the items is not available, items can be prescreened using classical test theory methods such as the Mantel Haenzel test, or logistic discriminant function analysis, as used in this study. A third way to establish a set of anchor items is through the study design. Depending on the groups being compared, it is often possible to administer some items to both groups in the same format. For example, in a study comparing responses to phone and mail surveys, a subset of items could be administered by phone to all participants and used as anchor items. This same strategy can be used to evaluate translated assessments with bilingual respondents (e.g., Sireci & Berberoglu, 2000). Once a set of anchor items is established, DIF is identified among the study items within this approach on the basis of model-based likelihood ratio tests (Thissen, Steinberg, & Wainer, 1988, 1993; Wainer, Sireci, & Thissen, 1991). Examples of this type of application can be found in Thissen, Steinberg, and Gerrard (1986), Sireci and Berberoglu (2000), and Teresi, Kleinman, and Ocepek-Welikson (2000).

Although both the DIF index and model comparison approaches have advantages and disadvantages, a major barrier to wide application of the first approach is the unavailability of commercial software to equate the parameters and calculate the DIF statistics. In contrast, the model comparison approach can be performed using Multilog (Thissen, 1991). Another important difference lies in the identification of statistically significant DIF. Whereas many of the DIF indices have been shown to perform adequately in simulation studies (e.g., Raju et al., 1995), it has been suggested that DIF analyses using model-based likelihood ratio tests are more powerful and should be emphasized over the DIF index approaches (Teresi et al., 2000; Thissen et al., 1993; Wainer, 1995).

There has been a recent increase in the application of psychometric techniques originally developed for use in educational testing to other areas of psychology, particularly with respect to use of IRT and assessment of DIF (e.g., Collins et al., 2000; Donovan et al., 2000; Fraley, Waller, & Brennan, 2000; Hambleton, 2000; Hays, Morales, & Reise, 2000; Krueger & Finger, 2001; Morales, Reise, & Hays, 2000; Orlando, Sherbourne, & Thissen, 2000; Saliba, Orlando, Wenger, Hays, & Rubenstein, 2000; Teresi et al., 2000; Waller et al., 2000), but even within this family of techniques, approaches vary, as do the pros and cons of their use. The broad goal of the present study is to demonstrate the advantages of using an IRT-based model comparison approach to evaluating measurement equivalence. The example presented below uses data derived from a prospective study of the mental health consequences of exposure to community violence to identify and examine DIF in a Spanish version of the PCL–C.

Method

Participants

Data were collected as part of a larger longitudinal study of the mental health consequences of exposure to community violence. Participants were recruited from among those admitted to a large Level I trauma facility in a predominantly Latino community, for treatment of physical injuries stemming from community violence. Participants were young adults, between the ages of 18 and 35 years, who sustained injuries inflicted by a person other than a family member or a former intimate sexual partner. Participants were not approached to complete a short interview to screen for eligibility until they were judged medically capable of being interviewed, as determined by discussions with medical staff. Multiple attempts were made to monitor, screen, and interview persons who were either not initially available or who appeared to be cognitively impaired or insufficiently alert. A total of 584 persons could not be screened for eligibility, and 55 persons chose not to participate in a screening interview. The primary reason for failure to screen was discharge before an approach could be made. In all, approximately 60% of age-eligible persons (n = 653) were screened.

Of 653 persons screened, 423 were eligible for the study. Of these, 413 (98%) completed a face-to-face structured interview conducted by trained lay interviewers within several days of hospital admission (M = 11.73 days, SD = 15.14 days). Two eligible persons declined to participate; 5 participants chose to terminate the interview before completion. For the purposes of the current study, participants with impaired cognition at admission (n = 22), as evidenced by a Glascow coma score (GCS; Teasdale & Jennett, 1974) of less than 14 (range: 3–15), were excluded

from the analytic sample, to be consistent with other studies (Marshall & Orlando, 2002). GCS scores were obtained from computerized medical records.

Interview

All interviewers, the majority of whom were fully bilingual and bicultural, were provided with 6 full days of training before conducting interviews. Training topics included orientation to the study and medical center, general interviewing skills, item-by-item review of the interview, and mock interviews with immediate feedback provided by survey staff. No interviewer was allowed to conduct interviews until survey staff were satisfied as to their competence. Interviewers received group supervision on a weekly basis throughout the study.

The interviewer, in consultation with each respondent, determined the choice of English or Spanish administration; the decision was based on the language with which a given respondent was most facile. The interviews took approximately 60 min to complete and covered a broad range of topics. The Spanish version of the instrument was developed using the double-translation procedures described by Brislin (1970). A bilingual translator first translated each scale from English to Spanish, whereupon a second bilingual translator translated the scale back into English. Discrepancies or changes in meaning were reconciled by the two translators in collaboration with other bilingual translation team members who were not directly involved in either of the initial translations. No single translator was involved in more than one sequence of translations.

Measures

Symptoms of PTSD were assessed using the English or Spanish version of the PCL–C. This instrument contains 17 items corresponding to the three *Diagnostic and Statistical Manual of Mental Disorders* (DSM–IV; American Psychiatric Association, 1994) symptom clusters of reexperiencing, avoidance, and hyperarousal. Responses to each item are rated using a 5-point scale, ranging from 1 (*not at all*) to 5 (*extremely*), indicating the extent to which respondents had been bothered by that symptom in the past 7 days.

This instrument was chosen over other available measures of PTSD for several reasons. First, in a sample of Vietnam veterans, Weathers et al. (1993) demonstrated that the scale had solid psychometric properties (α = .97; test–retest reliability of .96; convergent validity with other PTSD symptom scales). Additionally, the scale had been widely used, with applications in a variety of samples including motor vehicle crash survivors (Blanchard, Jones-Alexander, Buckley, & Forneris, 1996), breast cancer survivors (Andrykowski, Cordova, Studts, & Miller, 1998; Cordova, Studts, Hann, Jacobsen, & Andrykowski, 2000), persons with severe mental illness (Mueser et al., 2001), patients in primary care settings (Asmundson et al., 2000; Stein, McQuaid, Pedrelli, Lenox, & McCahill, 2000), and survivors of bone marrow transplantation (Smith, Redd, DuHamel, Vickberg, & Ricketts, 1999). Finally, although the dimensionality of PTSD in general, and the PCL-C in particular, had been extensively studied and debated and factor analytic results had varied, there was compelling evidence for one higher order general factor among the PCL–C items (Asmundson et al., 2000, Cordova et al., 2000).

Data Analysis

Dimensionality and anchor items. IRT analysis requires that the collection of items can reasonably be considered to be reflective of one underlying dimension. Because the factor structure of the PCL–C had been a source of debate and because evidence existed for one higher order factor, we viewed formal testing of competing factor structures as potentially complex, unneccessary for the current purposes, and beyond the scope of this article. Instead, we ensured that a one-factor solution was an adequate representation of these data by conducting a confirmatory factor analysis of the 17-item scale.

Next, logistic discriminant function analyses (Miller & Spray, 1993; Swaminathan & Rogers, 1990) were used to identify suitable PCL–C items for matching the two language groups on overall PTSD symptoms. Using discriminant function analyses, group membership was predicted from the total score on the scale, the score on the item (item main effect), and the product of the total score and the item score (interaction effect). Changes in relative fit based on the inclusion of the item main effect and the interaction effect reflected the extent to which a given item performed differently from the overall scale with respect to group membership (Miller & Spray, 1993). Items whose main effect or interaction effect resulted in significant changes in fit (using α = .05) were flagged as having potential DIF and designated as study items, whereas items whose main and interaction effects resulted in nonsignificant changes in fit were selected as anchor items for the IRT analysis.

Item calibration. IRT models were estimated using the Multilog program (Thissen, 1991) to evaluate the equivalence of English and Spanish versions of the PCL–C. For the current analyses, Samejima's graded model was used (Samejima, 1969, 1997). Within this application, the graded response model estimates a slope parameter (*a*) and four location parameters (*b*) for each PCL–C item. The magnitude of the slope parameter reflects the degree to which the item is related to the underlying construct being measured. The location parameters reflect the spacing of the item responses along the latent construct continuum (Hambleton & Swaminathan, 1985). A given item can exhibit DIF with respect to the slope parameter, indicating that the relationship between the item and the underlying construct is stronger in one group than in another (nonuniform DIF). DIF can also be manifested with respect to location parameters, demonstrating that the difficulty of the item varies as a function of group membership (uniform DIF).

DIF detection. To implement DIF analysis within Multilog, data are arranged so that there are common anchor items for the reference and focal groups. Each studied item is represented as two separate items, one item holding actual responses for the reference group and missing values for the focal group and the other item holding actual responses for the focal group and missing values for the reference group. In this instance, English speakers were established as the reference group against which to compare Spanish-speaking respondents. The change in model fit (as measured by −2*log-likelihood) between a model in which the parameters of the studied item were allowed to differ between the groups and a model in which these parameters were constrained to be equivalent is distributed as chi-square with degrees of freedom equal to the number of item parameters; in this case, there are 5 *df*s associated with each item (4 location parameters and 1 slope parameter). A chi-square difference test was used to evaluate the degree of DIF in each of the studied items (Thissen, Steinberg, Pyszczynski, & Greenberg, 1983). If the 5-*df* test was significant, we also evaluated the 1-*df* test for nonuniform DIF (slope parameter only is constrained to be equal for the two groups). In all IRT DIF analyses, we used a *p* value of .01, as recommended by Teresi et al. (2000).

Interpretation. Plots created on the basis of the IRT parameterization were used as diagnostic tools to understand the nature and impact of identified DIF. Item boundary response functions (BRFs) represent each item with (in this application) four curves tracing the probability of scoring at or above Response Categories 2 through 5, respectively. The often-plotted traceline representing a positive response to a 0/1 dichotomous item is actually a BRF, that is, the boundary between a positive and negative response, or the probability of scoring 1. By displaying a separate set of BRF curves for the two groups, these plots characterize the presence of DIF in each response category (Collins et al., 2000). Item response functions (IRFs) collapse across categories, tracing the expected score (from 1 to 5) on a given item across the PTSD symptom-severity continuum. Separate IRFs for the two groups illustrate the impact of DIF at the item level (Collins et al., 2000; Donovan et al., 2000). Finally, separate test response functions (TRFs) are used to display the impact of DIF at the level of the entire scale by collapsing across items (Donovan et al., 2000). To assess the practical impact of identified DIF, group mean differences between Span-

ish and English speakers were compared using both DIF-adjusted IRT scores and standardized PTSD scores from the observed data.

Results

Sample Description

The majority of the 384 study respondents were male (94%) and Latino (Latino 79%, African American 11%, Caucasian 3%, other 7%). The Latino participants were primarily of Mexican descent, with 77% indicating Mexico, 15% indicating Central America, and 8% indicating South America, Puerto Rico, or Cuba as the country–region they identify with most. Seventy-three percent of the sample ($n = 282$; 29% non-Latino, 71% Latino) completed an English-language version of the instrument, and 27% were administered a Spanish version ($n = 102$; 100% Latino). English speakers were slightly younger than Spanish-speaking participants, with an average age of 24.5 years ($SD = 6.1$) compared with 26.6 years ($SD = 5.7$), and tended to have more education: Forty-eight percent of English-speaking participants, compared with only 25% of the Spanish speakers, had completed high school or its equivalent. The two language groups also differed with respect to marital status. Among the English speakers, 67% had never been married, 25% were either married or cohabiting, and 8% were separated or divorced. The comparable percentages for the Spanish-speaking sample were 47%, 42%, and 11%, respectively. Finally, English speakers were more likely to have sustained injuries stemming from gunshots (61%), as opposed to other penetrating or blunt objects (39%), whereas the reverse was true for the Spanish speakers (43% and 57%, respectively).

Scale Description and Dimensionality

The means and standard deviations of the 17 PTSD items, as well as brief item content, are displayed in Table 1 separately for the English and Spanish speakers. Values for the total score ranged from 1 to 5, reflecting the entire range of PTSD symptom severity. The total score means, not significantly different for the two groups, $t(382) = 0.40$, $p = .69$, indicate that on average, these participants experienced mild to moderate symptom severity. Results of the confirmatory factor analysis suggested that a one-factor solution adequately represented the data. Standard model fit measures (Bentler, 1990; Bentler & Bonett, 1980; Browne & Cudeck, 1989; Steiger & Lind, 1980) were within acceptable ranges, $\chi^2(119, N = 386) = 477.57$ (comparative fit index = 0.81, Tucker–Lewis index = 0.78, root-mean-square error of approximation = 0.09), implying sufficient unidimensionality for the purposes of this investigation.

Table 1
Means and Standard Deviations of PCL–C Items According to Language of Administration

	English		Spanish	
Abbreviated item content	*M*	*SD*	*M*	*SD*
1. Intrusive recollections	2.92	1.30	2.85	1.45
2. Distressing dreams	2.33	1.15	2.09	1.41
3. Reliving–flashbacks	1.85	1.18	1.83	1.23
4. Distress toward cues	2.49	1.33	2.98	1.58
5. Reactivity toward cues	2.34	1.40	2.23	1.45
6. Avoid thoughts–feelings	2.17	1.24	2.26	1.34
7. Avoid people–places	2.30	1.14	2.14	1.49
8. Amnesia	2.09	1.24	1.80	1.28
9. Diminished interest	2.19	1.40	2.35	1.61
10. Detached from others	2.34	1.43	2.21	1.51
11. Restricted range of affect	1.86	1.13	1.70	1.27
12. Foreshortened future	1.94	1.23	2.20	1.45
13. Disturbed sleep	2.36	1.36	2.52	1.53
14. Anger outbursts	1.88	1.13	2.34	1.38
15. Poor concentration	1.83	0.98	1.93	1.22
16. Hypervigilance	2.81	1.30	2.95	1.54
17. Exaggerated startle	2.31	1.20	2.27	1.40
Total	2.24	0.72	2.28	0.84

Note. PCL–C = PTSD Symptom Checklist—Civilian Version.

Table 2
Results of Discriminant Function Analysis and Item Designation

	Main effect		Interaction effect		
Item	χ^2	*p*	χ^2	*p*	Designation
1	0.92	.34	1.24	.27	Anchor
2	4.69	.03	4.92	.03	Study
3	0.23	.64	2.78	.10	Anchor
4	11.44	<.01	4.85	.03	Study
5	1.55	.21	0.85	.36	Anchor
6	2.10	.15	0.75	.39	Anchor
7	6.62	.01	11.13	<.01	Study
8	7.36	<.01	0.09	.76	Study
9	6.99	<.01	0.74	.39	Study
10	3.29	.07	2.90	.09	Anchor
11	4.93	.03	12.98	.00	Study
12	10.42	<.01	1.25	.26	Study
13	1.27	.26	0.99	.32	Anchor
14	12.97	.00	0.79	.37	Study
15	0.34	.56	4.15	.04	Study
16	3.76	.05	3.14	.08	Study
17	1.18	.28	3.05	.08	Anchor

Note. See text for discussion of how anchor and study items were determined. $N = 384$, $df = 1$ for all chi-square statistics.

Identifying Anchor and Study Items

Logistic discriminant analyses of the 17 PTSD items identified 10 items with suspected DIF (designated as study items for the IRT DIF analysis). Nine (2, 4, 7, 8, 9, 11, 12, 14, 15) displayed significant main or interaction effects (at $\alpha = .05$), and 1 (Item 16) was marginally significant (see Table 2). The 7 remaining items (1, 3, 5, 6, 10, 13, 17) were identified as anchor items and used in the IRT DIF analysis to establish a link between the two groups.

IRT DIF Analyses

For each studied item, the fit of a model constraining the item parameters to equality between the two language groups was compared with a model allowing the parameters to be estimated freely for the two groups. Six of the 10 studied items showed significant DIF according to the 5-*df* chi-square difference test between these two models (see Table 3). Although one item (11) showed marginal slope DIF, it did not reach our a priori significance level, so was not examined further. A final model estimation allowing for uniform DIF in Items 2, 4, 7, 8, 11, and 16 yielded the parameter estimates in Table 4. In this model, the overall PCL–C

Table 3
Results of IRT DIF Analyses for 10 Study Items

Item	$\Delta\chi^2(5)$	p (no item DIF)	$\Delta\chi^2(1)$	p (no slope DIF)
2	35.0	<.01	0.1	.75
4	34.0	<.01	0.2	.66
7	40.0	<.01	0.2	.66
8	18.8	<.01	3.2	.07
9	8.2	.15	—	—
11	21.0	<.01	4.6	.03
12	6.9	.23	—	—
14	15.0	.01	—	—
15	7.0	.22	—	—
16	28.1	<.01	0	1

Note. IRT = item response theory; DIF = differential item functioning. Dashes indicate tests not conducted. $N = 384$ for all chi-square statistics.

score for the Spanish speakers was −0.06 relative to a 0.0 mean for the English speakers.

As shown in Table 4, no items displayed cross-group DIF with respect to slope values. Thus, interpretion of the item slope parameters applies to both the English- and Spanish-language versions of the scale. With the exception of the low slope value of 0.82 for Item 8, "amnesia," the slopes for the remaining 16 items were all above 1.0, ranging from 1.15 to 2.03, indicating that the majority of the items have a reasonably strong relationship to the underlying construct of PTSD symptom severity. Somewhat surprisingly, of the 17 items, Item 15, "poor concentration," had the highest slope and thus was the most salient single indicator of PTSD symptom severity.

As a group, the 10 non-DIF items covered a fairly wide range of the PTSD symptom-severity construct, with location parameters ranging from −1.04 for b_1 of Item 1 to 2.36 for b_4 of Item 3. However, 1 of the DIF items, Item 16, had b_1 parameters lower than −1.04 (−1.14 in the English-speaking sample and −1.26 in the Spanish speaking sample), and 5 of the 6 DIF items had b_4 values greater than 2.36 in the Spanish-speaking sample. This discrepancy implies that the coverage of the PTSD symptom-severity construct would be compromised, especially at the high end of the scale, if the 6 DIF items were summarily omitted from analyses.

Evaluating DIF Impact

Figure 1 contains the BRFs for each of the six DIF items. In these plots, dotted lines represent the Spanish speakers and solid lines represent the English speakers. In the top left panel of Figure 1, the first curve (moving from left to right) is the Spanish-speaking probability of scoring 2 or higher on Item 2, "distressing dreams," and the last curve is the Spanish-speaking probability of scoring 5 on this item. All of the response functions for the English speakers (solid lines) are contained within these two curves. This observation implies that with respect to this item, the extreme response categories are considered less extreme by the English speakers than by their Spanish-speaking counterparts with equal levels of PTSD symptom severity. To illustrate, a Spanish speaker with a PTSD symptom-severity score of −2 on the IRT scale has a probability of approximately .30 of endorsing Category 2 or higher, whereas an English-speaking respondent with the same level of severity has a probability of only about .05 to score higher than 1.

Items 7 and 11 (Panels 3 and 5 in Figure 1) show a very similar pattern for the two groups in that all four of the English BRFs are completely contained within the first and last Spanish BRFs.

Table 4
Final Item Parameter Estimates Allowing for Uniform DIF in Six Items

Item	a	b_1	b_2	b_3	b_4
1	1.81 (0.18)	−1.04 (0.14)	−0.11 (0.10)	0.26 (0.09)	1.40 (0.16)
2—English	1.49 (0.12)	0.09 (0.13)	0.80 (0.14)	1.10 (0.15)	1.96 (0.21)
2—Spanish	1.49 (0.12)	−0.97 (0.24)	0.52 (0.22)	1.16 (0.25)	2.55 (0.50)
3	1.64 (0.20)	0.26 (0.10)	1.08 (0.14)	1.42 (0.16)	2.36 (0.29)
4—English	1.72 (0.13)	−0.83 (0.13)	−0.16 (0.11)	0.22 (0.11)	0.89 (0.13)
4—Spanish	1.72 (0.13)	−0.84 (0.22)	0.44 (0.20)	0.65 (0.20)	1.86 (0.30)
5	1.89 (0.19)	−0.16 (0.10)	0.50 (0.09)	0.82 (0.11)	1.58 (0.17)
6	1.51 (0.18)	−0.38 (0.12)	0.61 (0.12)	1.09 (0.15)	2.26 (0.29)
7—English	1.60 (0.13)	0.16 (0.12)	0.69 (0.13)	0.99 (0.14)	1.66 (0.18)
7—Spanish	1.60 (0.13)	−0.79 (0.24)	0.51 (0.20)	1.19 (0.22)	2.55 (0.54)
8—English	0.82 (0.08)	0.72 (0.21)	1.80 (0.26)	2.21 (0.28)	3.41 (0.42)
8—Spanish	0.82 (0.08)	−0.44 (0.37)	1.27 (0.39)	2.02 (0.46)	3.60 (0.74)
9	1.15 (0.17)	−0.04 (0.14)	0.67 (0.16)	0.91 (0.19)	1.73 (0.29)
10	1.40 (0.18)	−0.09 (0.12)	0.67 (0.14)	0.96 (0.16)	1.69 (0.23)
11—English	1.50 (0.14)	0.73 (0.14)	1.29 (0.16)	1.57 (0.19)	2.14 (0.25)
11—Spanish	1.50 (0.14)	0.06 (0.19)	1.02 (0.25)	1.56 (0.28)	2.67 (0.56)
12	1.48 (0.18)	−0.01 (0.11)	0.74 (0.13)	1.07 (0.16)	1.91 (0.25)
13	1.06 (0.15)	−0.59 (0.17)	0.52 (0.16)	0.86 (0.19)	1.98 (0.31)
14	1.32 (0.16)	−0.38 (0.13)	0.68 (0.14)	1.22 (0.18)	2.24 (0.30)
15	2.03 (0.21)	−0.03 (0.09)	0.96 (0.11)	1.35 (0.14)	2.19 (0.24)
16—English	1.16 (0.09)	−1.14 (0.19)	−0.05 (0.16)	0.31 (0.15)	1.29 (0.18)
16—Spanish	1.16 (0.09)	−1.26 (0.31)	−0.34 (0.25)	0.27 (0.25)	2.81 (0.59)
17	1.74 (0.18)	−0.42 (0.11)	0.63 (0.11)	0.99 (0.13)	1.82 (0.20)

Note. Standard errors are given in parentheses. Separate English- and Spanish-language parameter estimates are presented for the six nonequivalent items. DIF = differential item functioning. a indicates slope parameter; bs refer to location parameters (please see text for discussion).

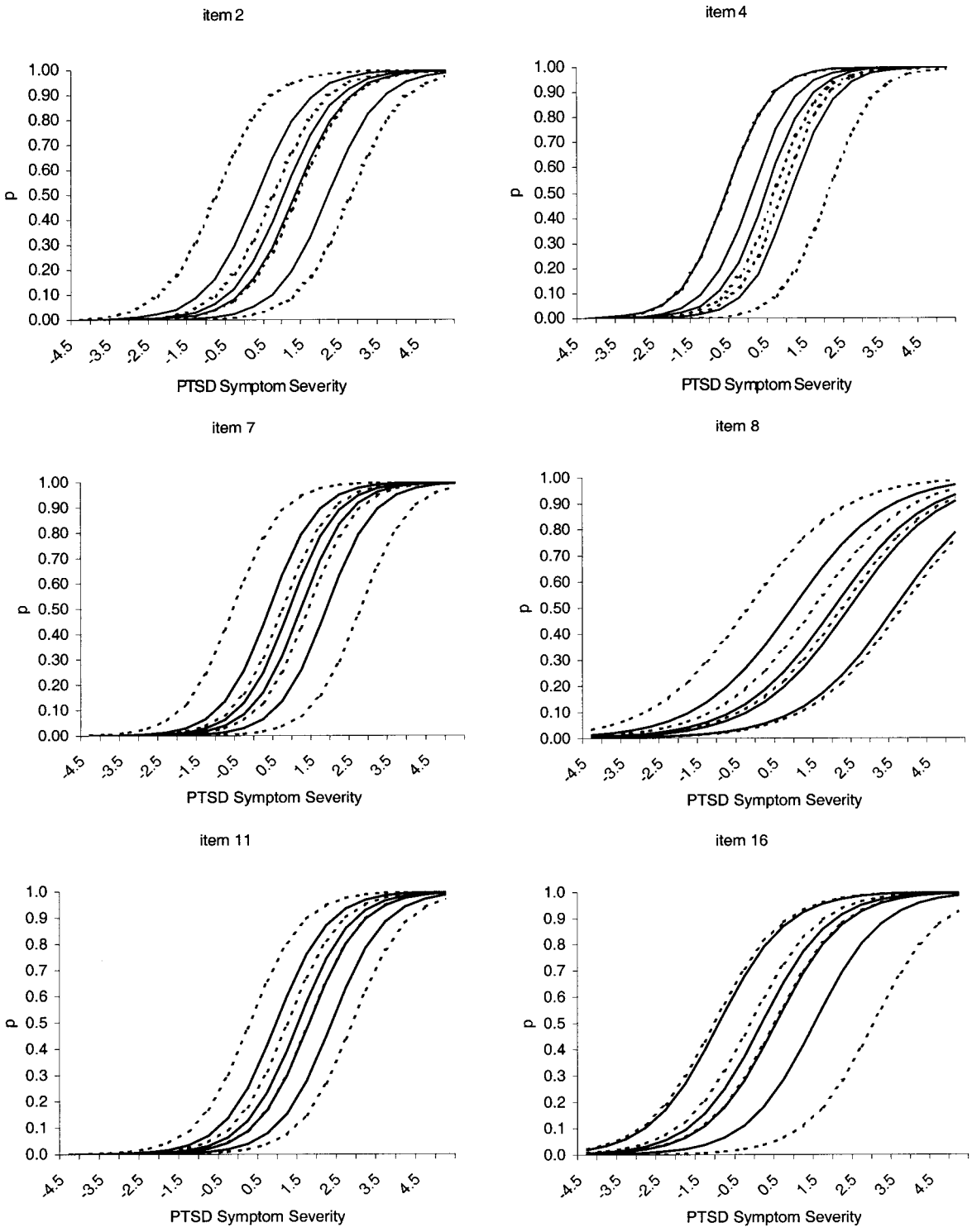

Figure 1. Boundary response functions for six differential item functioning items. Solid lines are English speakers; dotted lines are Spanish speakers. PTSD = posttraumatic stress disorder.

Items 8 and 16 (Panels 4 and 6) also have this quality, although the pattern is not as distinct. Finally, whereas the first boundary curve of Item 4 is fully coincident for the two groups, the Spanish speakers tend to endorse higher categories for this item than do their English-speaking counterparts with equal levels of PTSD symptom severity.

Inspection of the BRFs indicates that the two language groups are not using the response categories in the same fashion for these

six DIF items. It is not clear from these plots, however, what this finding means in terms of the two groups' total scores on these items. For insight into those differences, we look to the IRFs in Figure 2. The curves in these plots trace the expected item score associated with increasing levels of PTSD symptom severity. As with the BRFs, Items 2, 7, and 11 have similar patterns: For severity values at the low end of the continuum, Spanish speakers are likely to score higher on these items, but at the high end, the opposite is true. Item 16 also shows this pattern to a degree, although the difference at the low end is slight. In contrast, Spanish speakers score consistently lower on Item 4 and higher on Item 8 than do their English-speaking counterparts.

The plot in Figure 3 is a partial TRF composed of the six DIF items. This plot was created by summing the six IRFs in Figure 2 to get an indication of the total impact of the DIF in these items collectively. As can be seen in Figure 3, Spanish speakers tend to score slightly lower on these six items than severity-comparable English speakers at the low end of the PTSD symptom-severity scale, whereas English speakers tend to score slightly lower than severity-comparable Spanish speakers at the high end of the scale. Given that the effect of the DIF does not consistently favor one group over the other and that the distance between the curves for the two language groups looks negligible in Figure 3, it appears probable that the DIF in these six items will not have a significant impact on the overall PTSD symptom-severity scale scores.

As a practical evaluation of the impact of DIF, IRT scores that account for the DIF in the six items were compared with standardized PTSD symptom-severity scores from the observed data. The two sets of scores were very highly correlated ($r = .97$ for the full sample; $r = .97$ among the English speakers; $r = .97$ among the Spanish speakers). Generally speaking, the presence of DIF in a scale most directly influences estimates of group mean differences. In this case, however, the mean difference between the English and Spanish speakers was not greatly affected by accounting for the DIF. In fact, the difference between the two groups was small and nonsignificant, with English speakers scoring slightly higher than the Spanish-speaking group, regardless of whether the difference was calculated using DIF-adjusted IRT scores or standardized observed scores, $M = 0.06$, $t(384) = 0.53$, $p = .60$, and $M = 0.05$, $t(382) = 0.40$, $p = .69$, respectively.

Discussion

In this study we examined the comparability of English- and Spanish-language versions of the PCL–C using data derived from participants in a prospective study of the mental health consequences of exposure to community violence. Our results indicate that the Spanish translation of the PCL–C is not entirely equivalent to the English version. Although all items were equivalent with respect to slope estimates, 6 of the 17 PCL–C items exhibited DIF in location parameters. The lack of nonuniform DIF between the two language versions indicates that analogous items across the two groups are equally related to the underlying PTSD symptom-severity construct. The presence of uniform DIF in 6 items, however, indicates that the response categories for these items do not reflect the same degree of PTSD symptom severity across the two language versions. Thus, mean comparisons between language groups for these 6 items would be an inadequate reflection of true item-level group differences. A closer evaluation of the nature of the DIF and its impact revealed, however, that the problem was negligible at the level of the PCL–C scale. The combined TRF of the 6 biased items revealed that the bias essentially cancelled itself out at the aggregate level. Additionally, sensitivity analyses revealed that PTSD symptom-severity scores that accounted for DIF were highly correlated with standardized scores from the observed data, and conclusions about mean differences between language groups were not affected by DIF adjustment.

From a practical standpoint, the observed item bias in the Spanish version of the PCL–C is not particularly problematic for the purposes of the current program of research on the mental health consequences of community violence. Specifically, although it would not be appropriate to aggregate across English- and Spanish-language versions of the PCL–C when conducting analyses involving individual items, it appears reasonable to assume equivalence for scale-level analyses, particularly if sensitivity analyses are conducted to ensure that accounting for the observed DIF does not influence study findings.

Conclusions regarding the use of this translated assessment instrument in future studies are less definitive. When one or more DIF items are identified among a large item pool, as is often the case with educational tests, the items can be deleted without affecting overall measurement precision (because they can be replaced with equally efficient non-DIF items). However, in the measurement of psychological constructs, researchers commonly use established scales with psychometric properties that are vulnerable to a loss of even one item. This is particularly true for short scales and scales that are developed to reflect specific *DSM–IV* criteria. Ideally then, one would hope to use the information from a DIF analysis of a psychological scale to guide modifications aimed at eliminating the DIF in the item rather than eliminating the item itself.

The pattern of the BRFs for the six DIF items in this analysis imply that the English and Spanish speakers were not using the response scale comparably. On the whole for these 6 items, English speakers were more likely to endorse the highest response option, *extremely*, and the lowest response option, *not at all*, than Spanish speakers with equal PTSD symptom severity. A review of item content revealed no intuitive explanations for the observed differences between these 6 items and the remaining 11. Although our analyses and diagnostic plots allow us to describe the nature of the DIF, discernment of the source or likely cause of the DIF is difficult without input from translators and respondents. A necessary next step in the modification of the Spanish version of the PCL–C involves gathering this type of feedback, perhaps using a focus group format. Fortunately, the identified DIF in the Spanish version of the PCL–C does not have a great impact on scale scores, so modifying the instrument is not imperative for the purposes of future waves of this study. Similarly, researchers who wish to use this Spanish translation of the PCL–C in other studies may do so without great concern provided that sensitivity analyses are conducted to evaluate the impact of potential bias on study findings.

There are a number of limitations to consider when interpreting these results. Specifically, the results are generalizable only to the population from which the sample was drawn. In this instance, we focused on primarily Latino men who were recent survivors of community violence. Thus, scale equivalence of the Spanish and English versions of the PCL–C should be replicated in other samples experiencing PTSD symptoms after exposure to different traumatic events. Additionally, although the inclusion of non-Latinos in the English-speaking sample allowed us to conclude

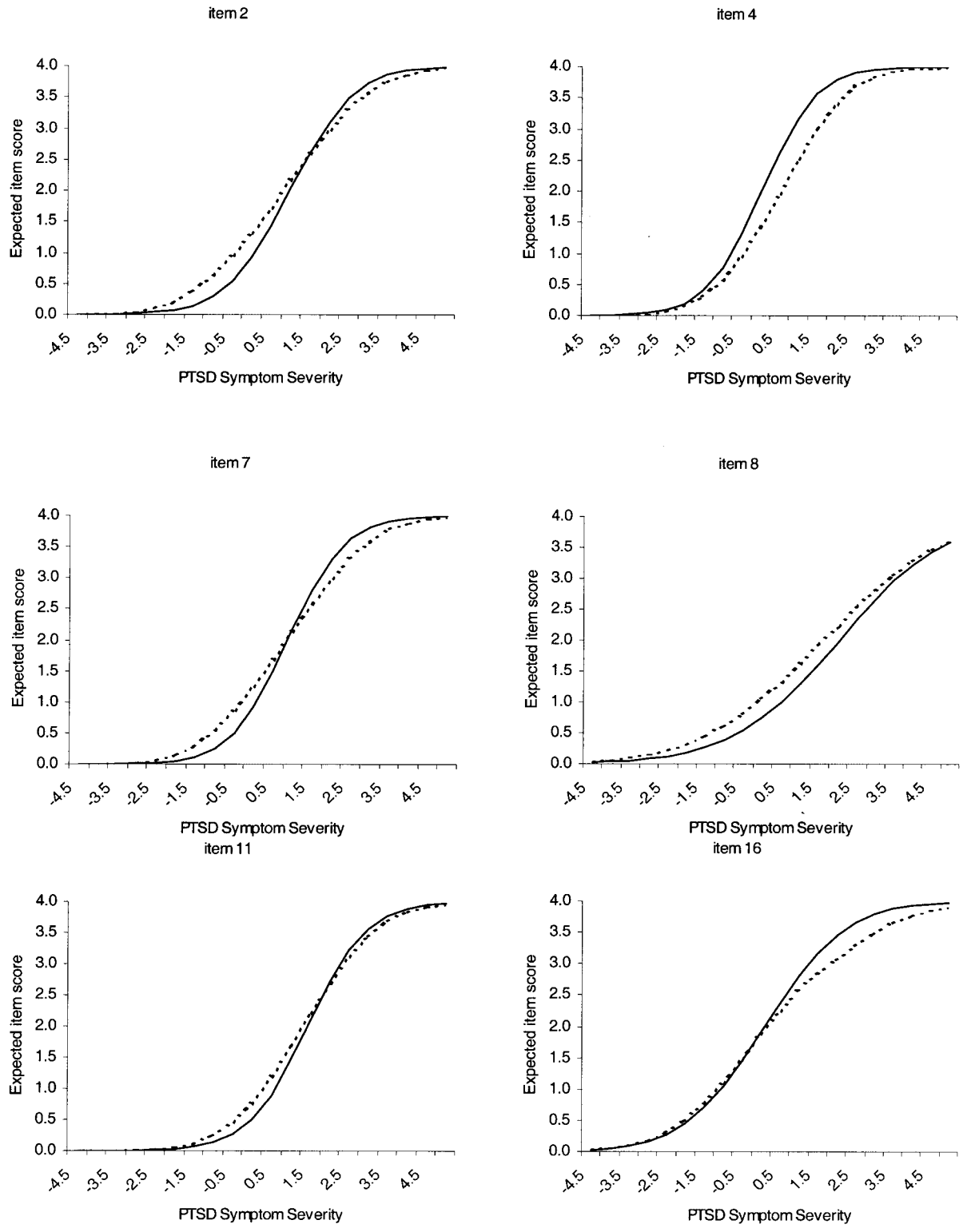

Figure 2. Item response functions for six differential item functioning items. Solid lines are English speakers; dotted lines are Spanish speakers. PTSD = posttraumatic stress disorder.

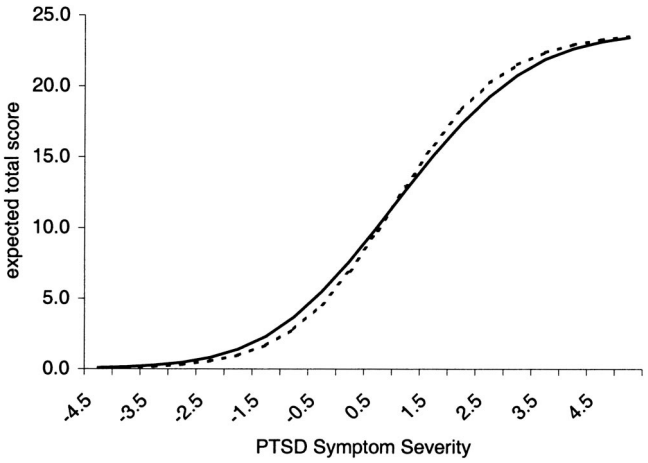

Figure 3. Test response function for six differential item functioning items. Solid line represents English speakers; dotted line represents Spanish speakers. PTSD = posttraumatic stress disorder.

that the two language versions can be combined in subsequent analyses of these data, it posed a slight confound with respect to interpreting the cause of the identified DIF. Specifically, it is not clear whether and to what extent the cultural imbalance in the two groups contributed to the DIF. However, the majority (71%) of respondents in the English-speaking sample were Latino, suggesting that the observed DIF was due to the translation. Similarly, we found that the two language groups differed significantly on a number of demographic characteristics. Although these differences are an unlikely source of the cause of identified DIF in this study, we cannot rule out that possibility. Future evaluations of the equivalence of the Spanish- and English-language versions of the PCL–C could avoid some of these confounds by including bilingual respondents in their study design. Finally, although this sample size is comparable to that observed in other IRT applications (e.g., Ellis, 1989; Thissen et al., 1986), it is smaller than would be optimal for obtaining precise IRT parameter estimates, thus limiting our ability to detect differences where they exist. However, it is evident that sample size did not appear to compromise our ability to detect DIF in this study.

In summary, translated assessment instruments are becoming increasingly necessary as the ethnic composition of our nation becomes more complex. It is essential to evaluate the measurement equivalence of such translated instruments, and IRT is a powerful methodology for this purpose. Results from analyses of the type described in this article can provide valuable diagnostic information as well as the ability to generate DIF-adjusted scores, enhancing the evaluation of the impact of nonequivalence on subsequent analyses.

References

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

Andrykowski, M. A., Cordova, M. J., Studts, J. L., & Miller, T. W. (1998). Posttraumatic stress disorder after treatment for breast cancer: Prevalence of diagnosis and use of the PTSD Checklist—Civilian Version (PCL–C) as a screening instrument. *Journal of Consulting and Clinical Psychology, 66,* 586–590.

Asmundson, G. J., Frombach, I., McQuaid, J., Pedrelli, P., Lenox, R., & Stein, M. B. (2000). Dimensionality of posttraumatic stress symptoms: A confirmatory factor analysis of *DSM–IV* symptom clusters and other symptom models. *Behavioral Research Therapy, 38,* 203–214.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF–36. *Journal of Clinical Epidemiology, 51,* 1189–1202.

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD checklist (PCL). *Behavioral Research Therapy, 34,* 669–673.

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

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995, December). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451–461.

Cordova, M. J., Studts, J. L., Hann, D. M., Jacobsen, P. B., & Andrykowski, M. A. (2000). Symptom structure of PTSD following breast cancer. *Journal of Traumatic Stress, 13,* 301–319.

Donovan, M. A., Drasgow, F., & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85,* 305–313.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74,* 912–921.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Garcia-Preto, N. (1996). Latino families: An overview. In M. McGoldrick, J. Giordano, J. K. Pearce (Eds.), *Ethnicity and family therapy* (2nd ed., pp. 169–182). New York: Guilford Press.

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

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52,* 1115–1124.

Guthery, D., & Lowe, B. A. (1972). Translation problems in international marketing research. *Journal of International Business Studies, 14,* 81–87.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis [Comment]. *Medical Care, 38*(Suppl. 9), II60–II65.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(Suppl. 9), II28–II42.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hulin, C., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Hillsdale, NJ: Dow Jones-Irwin.

Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment, 13,* 140–151.

Marshall, G. N., & Jaycox, L. (2000). *Spanish Translation of the Posttraumatic Stress Disorder Checklist—Civilian Version*. Unpublished manuscript.

Marshall, G. N., & Orlando, M. (2002). Acculturation and peritraumatic dissociation in young adult Latino survivors of community violence. *Journal of Abnormal Psychology, 111,* 166–174.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107–122.

Morales, L. S., Reise, S. P., & Hays, R. D. (2000). Evaluating the equivalence of health care ratings by Whites and Hispanics. *Medical Care, 38,* 517–527.

Mueser, K. T., Rosenberg, S. D., Fox, L., Salyers, M. P., Ford, J. D., & Carty, P. (2001). Psychometric evaluation of trauma and posttraumatic stress disorder assessments in persons with severe mental illness. *Psychological Assessment, 13,* 110–117.

Negy, C., & Snyder, D. K. (2000). Reliability and equivalence of the Spanish translation of the Marital Satisfaction Inventory—Revised (MSI–R). *Psychological Assessment 12,* 425–430.

Norris, F. H., & Perilla, J. L. (1996). The revised Civilian Mississippi Scale for PTSD: Reliability, validity, and cross-language stability. *Journal of Traumatic Stress, 9,* 285–298.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment, 12,* 354–359.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995, December). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.

Saliba, D., Orlando, M., Wenger, N. S., Hays, R. D., & Rubenstein, L. Z. (2000). Identifying a short functional disability screen for older persons. *Journals of Gerontology: Medical Sciences, 55,* M750–M756.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Unpublished manuscript.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated–adapted items. *Applied Measurement in Education, 13,* 229–248.

Smith, M. Y., Redd, W., DuHamel, K., Vickberg, S. J., & Ricketts, P. (1999). Validation of the PTSD Checklist—Civilian Version in survivors of bone marrow transplantation. *Journal of Traumatic Stress, 12,* 485–499.

Steiger, J. H., & Lind, J. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stein, M. B., McQuaid, J. R., Pedrelli, P., Lenox, R., & McCahill, M. E. (2000). Posttraumatic stress disorder in the primary care medical setting. *General Hospital Psychiatry, 22,* 261–269.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Swaminathan, H., & Rogers, H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *Lancet, 2*(7872), 81–84.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19,* 1651–1683.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983, Spring). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*(2), 211–226.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–186.

Wainer, H., Sireci, S. G., & Thissen, D. (1991, Fall). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5,* 125–146.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993, October). *The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility*. Paper presented at the International Society for Traumatic Stress Studies, San Antonio, TX.

Weidmer, B., Brown, J., & Garcia, L. (1999). Translating the CAHPS 1.0 survey instruments into Spanish: Consumer assessment of health plans study. *Medical Care, 37*(Suppl. 3), MS89–MS96.

Wells, K. B. (1999). The design of Partners in Care: Evaluating the cost-effectiveness of improving care for depression in primary care. *Social Psychiatry and Psychiatric Epidemiology, 34,* 20–29.

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and problems of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of methods in cultural anthropology* (pp. 398–420). New York: Columbia University Press.

Received May 16, 2001
Accepted November 8, 2001 ■