Health Serv Outcomes Res Method
DOI 10.1007/s10742-009-0052-4

A non-parametric Bayesian diagnostic for detecting differential item functioning in IRT models

Mark E. Glickman · Pradipta Seal · Susan V. Eisen

Received: 1 July 2008 / Revised: 13 April 2009 / Accepted: 10 July 2009

Abstract Differential item functioning (DIF) in tests and multi-item surveys occurs when a lack of conditional independence exists between the response to one or more items and membership to a particular group, given equal levels of proficiency. We develop an approach to detecting DIF in the context of item response theory (IRT) models based on computing a diagnostic which is the posterior mean of a *p*-value. IRT models are fit in a Bayesian framework, and simulated proficiency parameters from the posterior distribution are retained. Monte Carlo estimates of the *p*-value diagnostic are then computed by comparing the fit of nonparametric regressions of item responses on simulated proficiency parameters and group membership. Some properties of our approach are examined through a simulation experiment. We apply our method to the analysis of responses from two separate studies to the BASIS-24, a widely used self-report mental health assessment instrument, to examine DIF between the English and Spanish-translated version of the survey.

Keywords Bayesian modeling · Conditional independence · Mental health outcome · Model diagnostics · Patient surveys

The views expressed in this article are those of the authors and do not necessarily reflect the views of the Department of Veterans Affairs.

M. E. Glickman · S. V. Eisen
Department of Health Policy and Management, Boston University School of Public Health, Boston, MA, USA

M. E. Glickman (✉) · S. V. Eisen
Center for Health Quality, Outcomes and Economics Research, a Veteran Administration Center of Excellence, Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA
e-mail: mg@bu.edu

P. Seal
Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Published online: 24 July 2009

1 Introduction

The assessment of differential item functioning (DIF) has become an integral part of determining the validity of standardized tests and multi-item surveys. In the context of tests, DIF occurs when people from different groups with the same ability have systematically different responses to specific test items. If, for example, a math test item has boys answering correctly more often than girls of equal ability because the subject of the item is on a topic more familiar to boys (e.g., sports), then the item is said to exhibit DIF and should be considered for modification or removal from the test. DIF of an item can therefore be understood as a lack of conditional independence between an item response and group membership (often gender or ethnicity) given the same latent ability or trait.

While differential item functioning has been applied most traditionally to educational tests, DIF studies are increasingly finding application to health surveys. The focus of this paper is on health surveys, so we will henceforth view a patient's health as the latent trait in a DIF analysis. In a recent paper, Teresi (2006) has provided a review of statistical issues of DIF in health applications. Perkins et al. (2006) have examined DIF for items in a widely used health status instrument by age, education, race and gender groupings, and found many items to exhibit DIF. Pagano and Gotay (2005) have shown the presence of DIF by ethnic groups in a quality of life survey for cancer patients. Cauffman and MacIntosh (2006), in a recent mental health application, examine DIF by gender and ethnic groups of incarcerated juveniles in an instrument designed to identify mental health problems. As more health-related applications involve the detection of DIF to establish the validity of health surveys, the more crucial the statistical underpinnings for DIF detection continues to be.

Various methods for detection of DIF have been proposed over the past 25 years. The most commonly used approach is based on a Mantel–Haenszel analysis of the relationship between item responses and group membership conditional on an observed measure of ability (Holland and Thayer 1988), usually, in the context of tests, the total number of correctly answered items. Another common approach to detect DIF is to use log-linear or logistic models, as described in Kok et al. (1985) and Swaminathan and Rogers (1990). Recognizing that these methods involve conditioning on a measured surrogate of a latent trait, DIF detection has been more recently formulated in the context of item response theory (IRT) models. The advantage to the IRT framework is that latent trait is explicitly modeled as an unknown parameter to be inferred in the fitting process. Thissen et al. (1993) provide an overview of a set of methods that rely on fitting IRT models and then examining lack-of-fit statistics to assess the presence of DIF. These methods, however, require estimation of the latent trait and other model parameters (e.g., through maximum likelihood) on a likelihood surface which can be relatively flat due to the IRT model being highly parameterized. In such cases, the estimated latent trait parameters can be unreliable measures, even when evaluating likelihood-based quantities, and diagnostics based on these estimates can lead to overly optimistic conclusions.

To account for the uncertainty inherent in model inferences, several authors have explored DIF detection in IRT models within a Bayesian framework. One Bayesian approach is to determine the marginal posterior distribution of model parameters indicative of DIF. Wainer et al. (2007, ch 14), for example, suggest fitting an IRT model with a parameterization in which health outcomes depend explicitly on group membership—inferences about a group membership coefficient will reveal whether DIF has been detected for that item. The other main alternative has been to fit a single Bayesian model, and then perform posterior predictive checks (Gelman et al. 1996) as a means to diagnose

item-level lack of model fit. Hoijtink (2001), for example, proposes examining the posterior predictive distribution of a standardized lack-of-fit statistic to compare against the statistic evaluated on the observed data. To examine person-specific fit diagnostics in IRT models, Glas and Meijer (2003) propose using posterior predictive checks on a variety of measures. Sinharay (2005) provides a more general discussion of posterior predictive checks for Bayesian IRT models beyond the context of DIF detection. These approaches have promise in allowing some freedom to choose a relevant lack-of-fit measure, but the motivation for choosing particular measures is not always compelling.

This paper proposes a diagnostic method for detecting DIF in a Bayesian IRT model that relies on examining the posterior distribution of an appropriately chosen measure. Our method shares similarities with the approach of posterior predictive checks in that we first fit a Bayesian IRT model to obtain the posterior distribution of all model parameters. We then construct a measure that directly addresses whether a conventional definition of differential item functioning has been satisfied, and subsequently summarize the posterior distribution of this measure. Allowing the measure to be a function of the latent health parameters permits the diagnostic to address both the uncertainty in model inferences and the increased flexibility to specify a measure that appropriately captures DIF. Our method, however, is not a posterior predictive check as our diagnostic is not averaged over the posterior predictive distribution.

The paper is organized as follows. We explain the construction of our DIF diagnostic in Sect. 2. The method is evaluated through a simulation analysis in Sect. 3. The approach is then applied to a study on detecting DIF in items between an English and Spanish version of a commonly used mental health survey, which is presented in Sect. 4. A discussion of the method and its limitations are outlined in Sect. 5.

2 A Bayesian method for detecting DIF

Let i index respondents ($i = 1,..., n$) and let j index items ($j = 1,..., J$) on a J-item health survey. Assuming each item has K possible choices, consider a univariate IRT model of the form

$$\mathrm{P}(Y_{ij} = k|\alpha_j, \beta_{jk}, \theta_i, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{x}_i) \tag{1}$$

for $k = 1,..., K$, where θ_i is the latent health trait for respondent i, $\boldsymbol{x}_i$ is a vector of r covariates for respondent i, β_{jk} (for $k = 1,..., K-1$) is a "difficulty" parameter for the k-th category of item j, $\alpha_j > 0$ is an item-specific discrimination parameter, $\boldsymbol{\gamma}$ is a vector of other model parameters (for example, "guessing" parameters in certain IRT models), and $\boldsymbol{\delta}$ are the effects of $\boldsymbol{x}_i$. Several common examples of IRT models include the two-parameter logistic model (Birnbaum 1968) for binary responses,

$$\text{logit } \mathrm{P}(Y_{ij} = 1|\alpha_j, \beta_j, \theta_i) = \alpha_j(\theta_i - \beta_j), \tag{2}$$

the three-parameter logistic model (Birnbaum 1968) for binary responses,

$$\mathrm{P}(Y_{ij} = 1|\alpha_j, \beta_j, \theta_i, \boldsymbol{\gamma}) = \gamma_j + (1 - \gamma_j)\left(\frac{\exp(\alpha_j[\theta_i - \beta_j])}{1 + \exp(\alpha_j[\theta_i - \beta_j])}\right), \tag{3}$$

the generalized partial credit model (Muraki 1992)

$$\mathrm{P}(Y_{ij} = k|\alpha_j, \beta_j, \theta_i, \boldsymbol{\gamma}) = \frac{\exp \sum_{\ell=0}^{k} \alpha_j[\theta_i - (\beta_j - \gamma_{\ell j})]}{\sum_{x=0}^{K} \exp \sum_{\ell=0}^{x} \alpha_j[\theta_i - (\beta_j - \gamma_{\ell j})]}, \tag{4}$$

or the ordinal response model of Samejima (1969),

$$\text{logit } \mathrm{P}(Y_{ij} \geq k|\alpha_j, \beta_{jk}, \theta_i) = \alpha_j(\theta_i - \beta_{jk}). \tag{5}$$

In each of these models, person-specific background variables (e.g., socio-demographic variables) can be included in a straightforward manner by substituting θ_i with $\tilde{\theta}_i - \boldsymbol{x}_i'\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ in this case are the linear effects of $\boldsymbol{x}_i$. Adjusting the ability parameter by background covariate information can help with the identification of DIF, as argued by Glas (2001). The $\tilde{\theta}_i$ can be interpreted as a measure of health for participant i adjusted for socio-demographic effects, so that by viewing $\tilde{\theta}$ as the summary health feature all study participants are measured relative to the same baseline.

Suppose the question of interest is to determine which of the J test items exhibits DIF depending on membership to a "focal" group versus a reference group. For example, if one wants to examine whether certain items on a multiple choice exam favor boys relative to girls, the girls would be the focal group and the boys would be the reference group. Let g_i be the binary indicator of the focal group membership for respondent i, that is

$$g_i = \begin{cases} 0 & \text{if respondent } i \text{ is in the reference group} \\ 1 & \text{if respondent } i \text{ is in the focal group.} \end{cases} \tag{6}$$

Formally, for an arbitrary respondent, DIF exists for item j if, for some k,

$$\mathrm{P}(Y_j = k|\theta, g = 0) \neq \mathrm{P}(Y_j = k|\theta, g = 1) \tag{7}$$

(Hoijtink 2001, Shealy and Stout 1993). When (7) is true, either conditional independence is violated, or the assumption of unidimensionality of θ does not hold (see, for example, Angoff 1982). If covariate information, $\boldsymbol{x}$, is given, then, in terms of $\tilde{\theta}$, (7) can be restated as

$$\mathrm{P}(Y_j = k|\tilde{\theta}, g = 0) \neq \mathrm{P}(Y_j = k|\tilde{\theta}, g = 1). \tag{8}$$

The method we develop is constructed to detect when (7) and (8) do not hold for the sample of respondents in a study. For the remainder of the discussion, we will assume that covariate information is available, and that DIF detection will involve the $\tilde{\theta}_i$.

Our approach to detecting DIF for item j involves two steps. First, we fit an IRT model that may adjust for covariates $\boldsymbol{x}$, but does not adjust for DIF group membership g, within a Bayesian framework via Markov chain Monte Carlo (MCMC) simulation from the posterior distribution, and retain simulated values from the marginal posterior distribution of the $\tilde{\theta}_i$. Second, based on the results of the fitted model, we check whether the Y_j are conditionally independent of g given $\tilde{\theta}$. More specifically, for each simulated vector of health parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_n)$, we calculate the p-value for a likelihood ratio test comparing a flexible, possibly non-parametric, regression model for Y_j as a function of $\tilde{\boldsymbol{\theta}}$ and g to a smaller model for Y_j only as a function of $\tilde{\boldsymbol{\theta}}$. We assume that the choice of flexible models results in a likelihood ratio test statistic that is asymptotically χ^2-distributed, following classical theory. The average of these p-values across the simulated vectors of $\tilde{\boldsymbol{\theta}}$ is a Monte Carlo estimate of the posterior mean p-value. Because each individual likelihood ratio statistic is constructed to have a p-value that is approximately uniform under the model that does not include g, the resulting posterior mean p-value is also

calibrated to be approximately uniform. The reason is that the likelihood ratio statistic is being applied to the comparison of two flexible regressions, making it irrelevant that each Monte Carlo draw of $\tilde{\boldsymbol{\theta}}$ is simulated from the marginal posterior distribution of the IRT model. This second step of the algorithm is applied repeatedly for each item in the test. We discuss these two separate steps of our approach in detail below.

Bayesian fitting of IRT models is becoming increasingly commonplace arguably due to the increased ease of implementation of the fitting algorithms. In a pair of papers, Patz and Junker (1999a, b) lay out a general approach for implementing an MCMC algorithm for posterior sampling in the context of general IRT models. Other recent examples of Bayesian IRT modeling include Bradlow et al. (1999), Janssen et al. (2000), Beguin and Glas (2001), Fox and Glas (2001), Johnson and Sinharay (2005), and Wainer et al. (2007, ch 14). Rather than determining analytically the conditional posterior distributions necessary for MCMC simulation, publicly available Bayesian software such as WinBUGS (Spiegelhalter et al. 2003) and OpenBUGS (Thomas et al. 2006) allows for straightforward implementation of many IRT models. Recent examples of the use of WinBUGS in fitting IRT models include May (2006) who uses WinBUGS to fit multilevel IRT models, and Kang and Cohen (2007) who use WinBUGS in comparing methods of fit to various IRT models.

In determining the Monte Carlo posterior mean p-value, we calculate for each simulated $\tilde{\boldsymbol{\theta}}$ the usual p-value for the likelihood ratio χ^2 test comparing a model predicting Y_j from both g and $\tilde{\boldsymbol{\theta}}$ to a model predicting Y_j from only $\tilde{\boldsymbol{\theta}}$. It is important that the IRT model does not adjust $\tilde{\boldsymbol{\theta}}$ for g because the "null hypothesized" relationship between Y_j and $\tilde{\boldsymbol{\theta}}$ in the likelihood ratio test should not already be conditional on g. More formally, let $Q(\boldsymbol{Y}_j, g, \tilde{\boldsymbol{\theta}}|\mathcal{M}_1, \mathcal{M}_2)$ be the p-value for the χ^2 likelihood ratio test comparing models $\mathcal{M}_1 \subset \mathcal{M}_2$. Then the posterior mean p-value is computed as

$$\begin{aligned} q(\boldsymbol{Y}_j, g|\mathcal{M}_1, \mathcal{M}_2) &= \int Q(\boldsymbol{Y}_j, g, \tilde{\boldsymbol{\theta}}|\mathcal{M}_1, \mathcal{M}_2) p(\tilde{\boldsymbol{\theta}}|\boldsymbol{Y}, \boldsymbol{x}) d\tilde{\boldsymbol{\theta}} \\ &\approx \frac{1}{M} \sum_{m=1}^{M} Q(\boldsymbol{Y}_j, g, \tilde{\boldsymbol{\theta}}^{(m)}|\mathcal{M}_1, \mathcal{M}_2) \end{aligned} \tag{9}$$

where $\tilde{\boldsymbol{\theta}}^{(m)}$ is the m-th saved MCMC draw ($m = 1,..., M$). The choice of models $\mathcal{M}_1$ and $\mathcal{M}_2$ should allow for flexible relationships between $\boldsymbol{Y}_j$ and the two predictors; for example, for binary Y_j, a non-parametric logistic regression as a function of $\tilde{\boldsymbol{\theta}}$ and g is sensible. For polytomous Y_j, non-parametric models of Yee and Wild (1996) would be appropriate. Small values of the posterior mean p-value, $q(\boldsymbol{Y}_j, g|\mathcal{M}_1, \mathcal{M}_2)$, in (9) indicate evidence that the relationship between Y_j and $\tilde{\boldsymbol{\theta}}$ depends on g.

Our approach can be contrasted with that of Wainer et al. (2007, ch 14) who also develop a method of diagnosing DIF in a Bayesian model. To identify whether item j evidences DIF, their approach is essentially equivalent to fitting a Bayesian IRT model in (2) in which the parameter β_j is replaced by $(\beta_j(1-g_i) + \beta_j^* g_i)$, where β_j and β_j^* are the item difficulty parameters for the reference and focal groups, respectively. Posterior inferences about $P(|\beta_j - \beta_j^*| > 0)$ provide evidence of DIF for item j. The authors note that their procedure is computationally intensive, as separate models need to be fit for each DIF analysis of an item. They suggest a screening procedure in which the Mantel–Haenszel test identifies candidate items for DIF study under the Bayesian procedure.

Our approach is also similar to that of Hoijtink (2001), but can also be contrasted in several respects. The approach of Hoijtink more closely follows Gelman et al. (1996) in that the diagnostic DIF statistic (which itself is a standardized fit measure) is a function of observables, and that the posterior predictive p-value is computed based on comparing the

statistic evaluated on the observed data to the posterior distribution of the statistic from MCMC posterior predictive simulations.

Our method, in contrast, has several features that make it an appealing alternative to these two. First, unlike the Wainer et al. method, our approach requires fitting only one IRT model rather than one model per item, so that the Bayesian model fitting computation is confined to one analysis. Second, while Wainer et al. consider an additive effect (on the logit scale) of g_i, and Hoijtink propose a fit measure based on a crude surrogate of the latent trait (namely, for detecting DIF on item j, the sum of the scores for all other items), our method recognizes the possibility of a more complicated relationship, for example an interaction between the $\hat{\theta}_i$ and g_i through a non-parametric relationship with Y_{ij}. Third, our method does not require specifying the focal and reference groups prior to fitting the IRT model. Once posterior simulations of the $\tilde{\theta}_i$ have been obtained, a number of DIF analyses can be performed depending on dichotomies of interest. Finally, our measure is self-calibrated to have an interpretation as following a uniform distribution, so that the extra computation usually needed to obtain a reference distribution in a posterior predictive check is unnecessary.

3 Examination of method via simulation

To evaluate our approach in detecting DIF, we performed a small simulation experiment. Because each IRT model fit with MCMC posterior simulation and subsequent posterior mean p-value calculation can be computationally prohibitive, our simulation analyses are limited and intend only to provide a modest study of how various factors influence the ability of our method to assess DIF.

We generated binary outcomes from the 2-parameter logit IRT model specified in (2). We varied three factors in the simulation experiment:

1. the number of respondents, N (set to 150, 300, or 900)
2. the number of items, J (set to either 10 or 30)
3. the fraction of items, F, generated to exhibit DIF (set to either 10% or 20%)

This resulted in a total of $3 \times 2 \times 2 = 12$ simulation conditions. Within each condition, we repeated the process 10 times of simulating data and then implementing our approach for detecting DIF.

For any individual set of simulated data, we generated the α_j from a log-normal distribution with $\log \alpha_j \sim N(0, 0.25^2)$, the β_j from $N(0, 0.5^2)$, and the θ_i from $N(0, 1)$. We simulated Bernoulli g_i with probability 0.5, and for the fraction F of test items assumed to exhibit DIF we generated Bernoulli Y_{ij} according to

$$\text{logit P}(Y_{ij} = 1|\alpha_j, \beta_j, \theta_i, \gamma) = \alpha_j(\theta_i - \beta_j) - g_i\gamma, \tag{10}$$

with effect size $\gamma = 1.0$; for all other items, the Y_{ij} were simulated directly from (2). The choice of $\gamma = 1.0$ corresponds to an odds ratio of $\exp(1.0) \approx 2.7$, which has been considered a medium effect size in logistic regression (see, for example, Rosenthal 1996).

Once the response data were generated, we then fit the 2-parameter logistic IRT model in (2) but without the g_i as part of the model specification. We assumed a prior distribution that factored into independent densities with components $\log \alpha_j \sim N(0, \sigma^2)$, $\beta_j \sim N(0, 100)$, and $\theta_i \sim N(0, 1)$; such a constraint on the θ_i has been used previously, as in Wainer et al. (2007, ch 14). We also assumed a uniform prior density on σ between 0 and 100; this

type of prior density for standard deviations in hierarchical models has been recommended by Gelman and Hill (2007). Each MCMC sampler, which was implemented in OpenBUGS (Thomas et al. 2006) called from within the R (R Development Core Team 2008) using the R2WinBUGS function, was run with two parallel chains consisting of a burn-in period of 2000 iterations, retaining a subsequent 1000 simulated sets of θ_i from each chain for DIF analysis. From initial exploration, 2000 iterations appeared to be a sufficient number for the sampler to converge. Then, for each j, and the vector of the θ_i from iteration m, we computed a likelihood ratio χ^2-based p-value comparing the fit of a smoothing spline model of Y_{ij} regressed on the simulated θ_i, and the fit of a smoothing spline model of the Y_{ij} regressed on the interaction of θ_i and g_i (essentially one smoothing spline for $g_i = 0$ and a second for $g_i = 1$). Determining the p-value for this comparison is described in Hastie and Tibshirani (1990), and implemented with the "gam" function in R. The average of the 2,000 p-values is the Monte Carlo estimate of the posterior mean p-value.

Summaries from the simulations appear in Table 1. For the 10 replications across each simulation condition, we examined the distribution of posterior mean p-values for items assumed to have DIF, true and false positive rates (for DIF items and non-DIF items, respectively) relative to a 0.05 significance level, and Bonferroni-adjusted true and false positive rates in which the significance level is set to 0.05/J. When N is 150, the distribution of posterior mean p-values for the DIF items with the assumed effect size stays moderately large for all values of J and F. The probability of DIF detection is close to 0.5 for a 0.05 significance level, and is unacceptably low for the Bonferroni-adjusted significance level. The FPRs remain generally lower than expected under the uniform p-values. In doubling the sample size to 300, the p-values decrease to the 0.05–0.10 range with $J = 10$ items, and even lower (around 0.03) when $J = 30$. The TPR is between 70% and 90% for a 0.05 p-value, but only as high as 50% for the Bonferroni-adjusted analyses. Again, the FPRs are roughly consistent with a 0.05 level. With $N = 900$, the p-values for

Table 1 Summaries of the simulation experiment

N	J	F	DIF p-values			True and false positive rates			
			Mean	10%	90%	TPR	FPR	TPR(*)	FPR(*)
150	10	0.1	0.1527	0.0120	0.3040	0.5000	0.0000	0.0000	0.0000
150	10	0.2	0.0874	0.0016	0.1955	0.5500	0.0125	0.2000	0.0000
150	30	0.1	0.1380	0.0018	0.3016	0.5172	0.0444	0.1034	0.0037
150	30	0.2	0.1537	0.0041	0.5423	0.4833	0.0542	0.0833	0.0042
300	10	0.1	0.0960	0.0000	0.2097	0.7000	0.0000	0.5000	0.0000
300	10	0.2	0.0565	0.0002	0.1891	0.7500	0.0125	0.4500	0.0125
300	30	0.1	0.0290	0.0001	0.0796	0.8667	0.0222	0.5333	0.0037
300	30	0.2	0.0290	0.0001	0.0807	0.8000	0.0417	0.2833	0.0042
900	10	0.1	0.0007	0.0000	0.0010	1.0000	0.0444	0.9000	0.0000
900	10	0.2	0.0001	0.0000	0.0001	1.0000	0.0625	1.0000	0.0125
900	30	0.1	0.0001	0.0000	0.0002	1.0000	0.0333	0.9667	0.0000
900	30	0.2	0.0023	0.0000	0.0001	0.9833	0.1208	0.9833	0.0042

For each simulation condition indexed by values of N, J and F, the sample mean, 10 and 90 percentile of the p-value distribution from 10 replications are reported for items assumed to have DIF. TPR and FPR are the proportion of significant p-values at the 0.05 level for items assumed to have DIF and those not assumed to have DIF, respectively. TPR(*) and FPR(*) are the 0.05-level Bonferroni-adjusted true and false positive rates

the DIF items are very small, and at least 90% power is achieved even with the Bonferroni-adjusted significance levels. The FPRs are low, but some of the p-values incorrectly indicate the presence of DIF, especially when the fraction of DIF items F is 0.2. This may be due to the θ_i being estimated incorrectly from the wrong model (where the g_i are omitted from the models). However, with the Bonferroni-adjusted significance levels, the magnitude of the FPRs are not problematic.

4 Application to a mental health survey

We applied our method to examine DIF between an English and Spanish version of the Behavior and Symptom Identification Scale (BASIS-24), a commonly used mental health self-report instrument, for two Latino cohorts enrolled in mental health or substance abuse programs. The original 32-item BASIS instrument was developed in 1984, and was designed to be used as a mental health status measure from a patient's perspective for the outcome of mental health treatment (Eisen et al. 1994). Eisen et al. (2004) developed a revised instrument, the BASIS-24, containing 24 items, which is the focus of the current study. Reliability and validity of the BASIS-24 among Latinos was verified in Eisen et al. (2006) for the English version of the instrument, and in Cortés et al. (2007) and Eisen et al. (2009) for the Spanish translation.

While the English and Spanish instruments have been separately validated, it is of interest to know whether individual items have different meaning due to the nuances of the translation process or to inherent differences between the English and Spanish languages. The data we used to investigate this question came from two separate studies. The first sample consisted of the subset of self-identified Latinos among a cohort of English-speaking inpatients and outpatients receiving mental health or substance abuse treatment at programs across the U.S. The BASIS-24 assessments were made at the start of the study, with data collected 2001–2002 (Eisen et al. 2006). A total of 370 BASIS-24 assessments were available for our study. The second sample consisted of Spanish-speaking self-identified Latinos who were given the Spanish adaptation of the BASIS-24. The Spanish assessments were conducted from 2004–2005 and resulted in a total of 594 patients from three regions of the U.S. (Eisen et al. 2009). Sample summaries of these two sets of patients appear in Table 2. The English cohort contains a greater proportion of outpatients, tends to be younger, has a slightly greater proportion of male patients, is somewhat more educated, and has greater proportion of substance abuse patients and lower proportion of patients with depressive and schizophrenia/schizoaffective disorders compared to the Spanish cohort. To account for the imbalance on these background characteristics, we incorporate these features in our IRT models.

The BASIS-24 instrument consists of 24 items on a 5-valued ordinal scale indicating the degree of difficulty (none, a little, moderate, quite a bit, extreme) or frequency of symptoms experienced in the past week. The list of items in English and Spanish appears in Table 3. The 24 items comprise six domains: depression/functioning, interpersonal relationships, self-harm, emotional lability, psychotic symptoms, and substance abuse. Prior to modeling, we inverted the scale of six items (items 4 through 9) so that higher-valued responses always indicated worse mental health. Sample means and 95% confidence intervals of the individual item scores, stratified by English versus Spanish, are displayed in Fig. 1. Generally, the mean scores by item tend to be close between English and Spanish cohorts, though, for some items (including items 10, 12, 15, 16, 17), patients in the Spanish

Table 2 Summaries of patient sample stratified by cohort

Sample characteristic	Spanish cohort ($n = 594$)	English cohort ($n = 370$)
Patient status		
Inpatient	48	30
Outpatient	52	70
Age (years)		
Age < 25	15	22
$25 \leq$ Age < 35	24	39
$35 \leq$ Age < 45	25	24
$45 \leq$ Age < 55	25	13
$55 \leq$ Age	10	3
Gender		
Male	44	50
Female	56	50
Educational level		
4th grade–8th grade	31	7
8th grade–12th grade	30	27
High school graduate	14	32
Some college	18	23
4-year college graduate	6	11
Not recorded	2	1
Primary diagnosis		
Schizophrenia/Schizoaffective disorder	22	11
Depressive disorder	43	20
Bipolar disorder	9	7
Alcohol/drug use order	6	24
Anxiety disorder and others	14	11
Not recorded	6	28

Values represent percentage out of the respective cohort

cohort reported worse mental health. This difference could be due to the higher proportion of inpatients in the Spanish sample.

We modeled the BASIS responses using Samejima's (1969) IRT model for ordinal outcomes, incorporating the covariate adjustment term as described in Sect. 2. Specifically, for $i = 1,..., 594 + 370 = 964$, $j = 1,..., 24$, and $k = 1,..., 5$, we assumed

$$\text{logit } \mathrm{P}(Y_{ij} \geq k | \alpha_j, \beta_{jk}, \tilde{\theta}_i, \boldsymbol{x}_i, \boldsymbol{\delta}) = \alpha_j(\tilde{\theta}_i - \boldsymbol{x}_i'\boldsymbol{\delta} - \beta_{jk}), \tag{11}$$

where Y_{ij} is the response by patient i to item j, $\tilde{\theta}_i$ is the covariate-adjusted health measure for patient i, α_j and β_{jk} are as defined in (5), and $\boldsymbol{x}_i'\boldsymbol{\delta}$ is the linear effect (on the logit scale) of patient status (inpatient vs. outpatient), age, gender, educational level, and primary diagnosis, as they are categorized on Table 2. A small fraction of patients had their educational level and primary diagnosis missing, so we assumed a priori that a missing category was uniformly distributed over the observed category levels (though model fitting would likely reveal non-uniform posterior inferences).

Table 3 BASIS-24 items in English and Spanish

English	Spanish
During the past week, how much difficulty did you have...	*Durante la semana pasada, ¿ Qué tan difícil fue para usted...*
1. Managing your day-to-day life?	1. hacerse cargo de su vida diaria?
2. Coping with problems in your life?	2. enfrentar los problemas de su vida?
3. Concentrating?	3. concentrarse?
During the past week, how much of the time did you...	*Durante la semana pasada, ¿ Con cuánta frecuencia...*
4. Get along with people in your family?	4. se llevó bien con sus familiares?
5. Get along with people outside your family?	5. se llevó bien con personas que no son familiares suyos?
6. Get along well in social situations?	6. se llevó bien in situaciones sociales?
7. Feel close to another person?	7. se sintió cercano(a)a alguna otra persona?
8. Feel like you had someone to turn to if you needed help?	8. sintió que tenía alguien con quien contar si necesitaba ayuda?
9. Feel confident in yourself?	9. se sintió de sí mismo(a)?
10. Feel sad or depressed?	10. se sintió triste o deprimido(a)?
11. Think about ending your life?	11. pensó en quitarse la vida?
12. Feel nervous?	12. sintió nervioso(a)?
During the past week, how often did you...	*Durante la semana pasada, ¿ Que tan a menudo...*
13. Have thoughts racing through your head?	13. pensó muchas cosas muy rápido y todas la vez?
14. Think you had special powers?	14. pensó que tenía poderes especiales que otras personas no tienen?
15. Hear voices or see things?	15. oyó voces o vio cosas que otras personas no oyeron o vieron?
16. Think people were watching you?	16. creyó que las personas lo/la estaban vigilando?
17. Think people were against you?	17. creyó que la gente estaba en contra suya?
18. Have mood swings?	18. tuvo cambios inesperados de ánimo?
19. Feel short-tempered?	19. se sintió irritable?
20. Think about hurting yourself?	20. pensó hacerse daño?
21. Did you have an urge to drink alcohol or take street drugs?	21. tuvo muchas ganas de tomar alcohol o de usar drogas?
22. Did anyone talk to you about your drinking or drug use?	22. alguien le dijo algo sobre su uso de alcohol o drogas?
23. Did you try to hide your drinking or drug use?	23. trató de esconder su uso de alcohol o drogas?
24. Did you have problems from your drinking or drug use?	24. tuvo problemas debido a su uso de alcohol o drogas?

Because our IRT model was highly parameterized, several modeling restrictions and simplifications were made before implementing the MCMC posterior sampler. First, as in the simulation analyses, we assumed exchangeable prior density components, $\tilde{\theta}_i \sim N(0,1)$. Secondly, to properly identify covariate effects and to avoid unnecessary correlations among the covariate parameters, all individual covariate effects were constrained to sum to 0. Furthermore, the conditional posterior distribution of the individual β_{jk} given the remaining parameters were constrained to be sampled from a range limited by the adjacent parameter values, $\beta_{j,k\text{-}1}$ and $\beta_{j,k+1}$. Diffuse but proper prior density components were assumed for all model parameters.

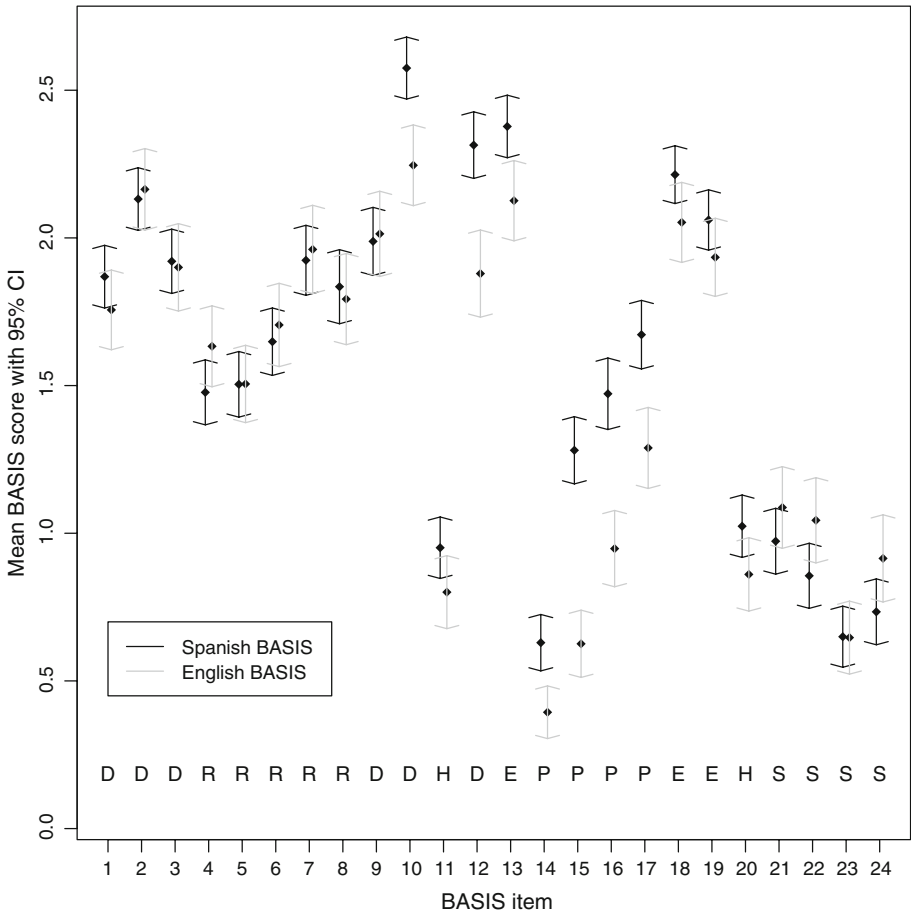

Fig. 1 Means and 95% confidence intervals for BASIS item scores, stratified by English/Spanish cohort. Items are labeled by categorization into six domains: "D", depression/functioning; "R", interpersonal relationships; "H", self-harm; "E", emotional lability; "P", psychotic symptoms; and "S", substance abuse

An MCMC sampler for the IRT model was implemented in OpenBUGS. Two parallel samplers were run for a burn-in period of 2000 iterations, after which the samplers were diagnosed to have converged through the examination of trace plots of various model parameters and through the examination of diagnostics such as the potential scale reduction statistic (Gelman and Rubin 1992). Simulated values of the $\tilde{\theta}_i$ were saved for the next 1,000 iterations in each chain, resulting in 2,000 simulated sets of parameter values. For the subsequent discussion, let $\tilde{\theta}_i^{(m)}$ denote the m-th iteration of $\tilde{\theta}_i, m = 1, \ldots, 2000$, from the MCMC sampler.

Letting $g_i = 1$ for the patients who were administered the Spanish version of the BASIS-24, and $g_i = 0$ for the English version, we carried out three analyses for each item j to compute posterior mean p-values to assess the lack of conditional independence of item responses and version of the BASIS-24 instrument given latent health measure $\tilde{\theta}$. First, for each m, treating the item responses as a 5-valued quantitative variable, we computed the likelihood ratio χ^2-based p-value comparing the fit of a smoothing spline model of Y_{ij}

regressed on the $\tilde{\theta}_i^{(m)}$, and the fit of a smoothing spline model of the Y_{ij} regressed on the interaction of $\tilde{\theta}_i^{(m)}$ and g_i as carried out in the simulation analyses of Sect. 3 using the "gam" function in R. The second analysis was the same as the first, except that Y_{ij} was modeled as a multinomial variable, the $\tilde{\theta}_i^{(m)}$ were discretized into five ordered categories of equal size, and the likelihood ratio was computed based on the fits of multinomial logit models. The third analysis also modeled the Y_{ij} as a multinomial variable, but did not discretize the $\tilde{\theta}_i^{(m)}$. A likelihood ratio p-value for this situation was constructed by fitting a multinomial logit model with a smoothing spline function of the $\tilde{\theta}_i^{(m)}$, and with a smoothing spline function of the interaction of $\tilde{\theta}_i^{(m)}$ and g_i. This modeling approach is described in Yee and Wild (1996), and has been implemented in the "vgam" function (Yee 2006) in R. In all three approaches, the average of the 2,000 p-values was the Monte Carlo estimate of the posterior mean p-values. The resulting values are displayed in the first three

Table 4 Results of the DIF analyses of the BASIS-24 responses

BASIS-24 Item	Posterior p-values			Likelihood p-values		
	Scenario A	Scenario B	Scenario C	Scenario A	Scenario B	Scenario C
1.	0.3211	0.2064	0.3054	0.1234	0.1132	0.1135
2.	0.0089	0.0191	0.0031	0.5613	0.0109	0.0044
3.	0.0523	0.1861	0.1724	0.0718	0.1866	0.3300
4.	0.0454	0.0144	0.0013	0.0006	0.0012	0.0003
5.	0.0062	0.3122	0.1024	0.0107	0.1981	0.0237
6.	0.0140	0.3386	0.2127	0.0003	0.0015	0.0327
7.	0.0005	0.0605	0.0101	0.0000	0.0053	0.0006
8.	0.0283	0.5317	0.3751	0.0875	0.8029	0.6692
9.	0.0011	0.0374	0.0198	0.0009	0.0772	0.0277
10.	0.3010	0.2636	0.1038	0.0702	0.0578	0.0754
11.	0.4968	0.6595	0.6353	0.0088	0.0420	0.2649
12.	0.0395	0.1104	0.0393	0.0504	0.0866	0.0872
13.	0.1563	0.2813	0.1734	0.1043	0.5369	0.2123
14.	0.1141	0.0618	0.0653	0.6637	0.1106	0.1017
15.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16.	0.0000	0.0162	0.0025	0.0000	0.2330	0.0282
17.	0.0030	0.0127	0.0012	0.0039	0.0007	0.0022
18.	0.1946	0.0695	0.0113	0.0444	0.0047	0.0019
19.	0.0001	0.0000	0.0000	0.0086	0.0000	0.0000
20.	0.0000	0.1063	0.0156	0.0000	0.0054	0.0331
21.	0.0034	0.1996	0.1351	0.0014	0.0600	0.0179
22.	0.2738	0.2567	0.2220	0.4978	0.2315	0.2890
23.	0.1414	0.3506	0.2861	0.8267	0.0897	0.1657
24.	0.0000	0.0068	0.0003	0.0000	0.0002	0.0000

The first three columns display posterior mean p-values for the Bayesian analyses based on the posterior draws of the $\tilde{\theta}$, and the latter three columns show the p-values resulting from likelihood analyses using the mean response of all but the item in question as the health measure. Scenario A treats the item response as quantitative and the health measure as quantitative; Scenario B treats the item response as multinomial and the health measure as categorical; and Scenario C treats the item response as multinomial and the health measure as quantitative. The boxed p-values are significant at the 0.05 level with a Bonferroni adjustment for each of the 24 items

columns of Table 4 (the corresponding methods above are labeled Scenarios A, B and C on the table).

In addition to the three analyses above, we performed three likelihood-based analyses that paralleled the Bayesian analyses. Our likelihood analyses for item j involved replacing $\tilde{\theta}_i^{(m)}$ with $\bar{Y}_{i(-j)} = \frac{1}{J-1}\sum_{\ell \neq j} Y_{i\ell}$ in each instance. Thus, each of our likelihood-based p-values was the result of comparing a model that regressed Y_{ij} on $\bar{Y}_{i(-j)}$ and g_i, assessing the significance of g_i. The use of $\bar{Y}_{i(-j)}$ as a proxy for the latent measure has been used conventionally, as in Junker (1993), Zhang and Stout (1999), and Hoijtink (2001). As our likelihood analyses parallel the Bayesian analyses, we examine the results of three sets of models depending on whether both Y_{ij} and $\bar{Y}_{i(-j)}$ are treated as quantitative, whether both are treated as categorical variables, or whether Y_{ij} is categorical while $\bar{Y}_{i(-j)}$ is quantitative. The results of these analyses are presented in the final three columns of Table 4.

The likelihood-based and posterior mean p-values in Table 4 reveal that the Bayesian diagnostic tends to be slightly more conservative than the likelihood-based diagnostic, as the latter tends to produce smaller values. Treating p-values that were significant at the 0.05 level, accounting for a Bonferroni adjustment of 24 items (that is, p-values that were less than 0.05/24 = 0.0021), as evidence of DIF between the English and Spanish versions of the BASIS-24, a greater number of items were flagged by the likelihood-based method. These p-values are highlighted on Table 4. Not surprisingly as well, Scenario B generally results in the largest p-values among the three modeling scenarios because both the Y_{ij} and the health measure are treated as categorical variables, while Scenario A tends to produce the most significant p-values as both the Y_{ij} and the health measure are modeled as quantitative variables. Scenario B of the likelihood analyses corresponds to the most common procedure involving log-linear models, and results in identifying six items exhibiting DIF. We suggest that Scenario C of the Bayesian approach, which models the Y_{ij} as multinomial and incorporates the effect of $\tilde{\theta}_i$ as a smoothing spline relationship, is the most consistent with modeling assumptions. This particular posterior mean p-value identifies five items as evidencing DIF, and these are a subset of the six identified in Scenario B of the likelihood analysis. It is interesting to note that the BASIS-24 item that is identified in the likelihood analysis to exhibit DIF but not in the Bayesian analysis (item 6) has markedly differently p-values.

Considering the five items exhibiting DIF, several points are noteworthy. Items 15 and 17 ("hear voices or see things," and "think people were watching you") are both part of the psychotic symptoms domain. The DIF found for some psychotic symptoms is consistent with other reports in the literature suggesting that psychotic symptoms such as hearing voices or seeing things may reflect Latino cultural or spiritual beliefs rather than signs and symptoms of psychotic disorders (Geltman et al. 2004; Guarnaccia et al. 1992; Vega et al. 2006). One item exhibiting DIF (item 19, feel short-tempered), proved especially difficult to translate. There is no Spanish equivalent to the English term "short-tempered." The closest approximation was the Spanish word "irritable," which translates to irritable in English. Consequently, DIF may have occurred due to the inadequacy of the translation of this term. Reasons for DIF on the remaining two items (getting along with people in your family and having problems from drinking or drug use) are unclear, as there appeared to be no difficulty with translation, and no obvious cultural influences on the understanding of these areas. Further research is needed to determine whether DIF on these items can be accounted for by other factors such as acculturation, education or other variables.

5 Discussion

The method described in this paper to detect DIF in multi-item health surveys is both a flexible and computationally feasible approach compared with alternative methods. Our method relies on fitting a single Bayesian IRT model and saving Monte Carlo simulated health parameters from the fit, followed by performing a separate analysis that examines whether the DIF grouping variable is predictive of survey responses beyond the health parameters. Each of these steps can be implemented using standard statistical software. An attractive feature of our approach is that it explicitly incorporates the uncertainty in the latent health measure in detecting DIF through repeated evaluations of the likelihood ratio p-value averaged over the Monte Carlo simulated vectors of $\tilde{\boldsymbol{\theta}}$.

An important difference between our approach and more conventional approaches is that, because we fit an IRT model before carrying out DIF diagnosis, inferences about the latent health status are formed using information from all item responses, not excluding the item about which DIF detection is being performed. This is mainly due to construction; our method allows for examining DIF on a variety of groupings after the IRT model has been fit. At worst, incorporating information from all responses in the IRT model might result in slightly more conservative inferences about DIF for each item, but this small loss in efficiency is offset by a gain in computational simplicity through the need to fit only one IRT model. However, based on the simulation analyses, it appears that a combination of larger data sets along with a large fraction of items with DIF can increase the false positive rate of DIF detection because the health parameters are not inferred correctly.

Another notable feature of our two-step approach is that the relationship of the response, Y_{ij}, and the latent health measure, $\tilde{\theta}_i$, in the IRT model is patently different from the more flexible relationship assumed in the posterior mean p-value computation. The reason for this approach is that detecting DIF is a diagnostic procedure that uses the $\tilde{\theta}_i$ as a proxy for latent health rather than specifically as an IRT model parameter, so that the approach to assess conditional independence between the response and DIF grouping can treat $\tilde{\theta}_i$ in a flexible relationship. In this manner, our approach has connections with the Mantel–Haenszel non-parametric approach.

Because our method separates model fitting and DIF assessment, many extensions to our approach are straightforward to implement. For example, assessing DIF as the comparison among more than two groups (i.e., treating g_i as a categorical variable with an arbitrary number of levels) poses no difficulties, as the likelihood ratio computation would simply incorporate g_i as a categorical variable appropriately. Differential test functioning, in which some or all items of a multi-item survey or test are combined as a weighted combination (or simply as an unweighted sum) to produce clinically meaningful survey summaries also pose no difficulties for our approach. After the IRT model is fit to the response data as usual, the likelihood ratio comparison of non-parametric regressions would then involve replacing individual items Y_j by subscale scores or entire survey scores, and posterior mean p-values would then be computed in the usual manner. Our method could also be extended to multi-dimensional IRT models (see Gardner et al. 2002, for a multidimensional extension of the Samejima model), in which θ_i is a vector-parameter; MCMC-simulated draws of the θ_i are retained, and the posterior mean p-value are computed as the comparison of the two non-parametric multiple regressions of the Y_{ij} on the θ_i alone and with the g_i.

Several limitations of our approach are worth noting. With great flexibility to choose a particular model to assess conditional independence (choice of categorizing variables, particular smoother for the $\tilde{\theta}_i$), the conclusions about items exhibiting DIF may depend

heavily on the choice. In the BASIS-24 analysis, Table 4 shows that treating the responses as quantitative usually yields much lower posterior mean p-values than the categorical response models. The two categorical response DIF analyses have a greater degree of agreement in the conclusions, but in some cases the p-values can be on the order of a factor of 10 apart, or higher (e.g., BASIS-24 items 4 and 17). Also, our method (as with most other IRT approaches) relies heavily on the IRT model being a reasonably correct representation of the data, and being properly specified (e.g., correctly incorporating covariate information, correct parameterization of discrimination and difficulty parameters, etc.). In particular, most DIF diagnostics, including ours, assume that when evaluating a specific item, other items are free of DIF. This is not ever likely to be the case, so a tacit assumption is that the number of items where DIF may be problematic is minimal. On the positive side, model misspecification will likely lead to more uncertain posterior inferences about the $\tilde{\theta}_i$, so that the diagnostic analyses using posterior samples will in turn lead to insufficient evidence of DIF. Thus our method is protective of false positives in the event that IRT models are inappropriately specified. But with an IRT model that has undergone appropriate model diagnosis and criticism, our method for detecting DIF is worthy of consideration.

Acknowledgments This research was supported by Grant R01 MH58240 from the National Institute of Mental Health and by the Veterans Administration Health Services Research and Development program.

References

Angoff, W.H.: Use of difficulty and discrimination indices for detecting item bias. In: Berk, R.A. (ed.) Handbook of Methods for Detecting Test Bias, pp. 96–116. Johns Hopkins University Press, Baltimore, MD (1982)

Beguin, A.A., Glas, C.A.W.: MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika **66**, 541–562 (2001)

Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (eds.) Statistical Theories of Mental Test Scores, pp. 396–479. Addison-Wesley, Reading, MA (1968)

Bradlow, E.T., Wainer, H., Wang, X.: A Bayesian random effects model for testlets. Psychometrika **64**, 153–168 (1999)

Cauffman, E., MacIntosh, R.: A Rasch differential item functioning analysis of the Massachusetts youth screening instrument. Educ. Psychol. Meas. **66**(3), 502–521 (2006)

Cortés, D.E., Gerena, M., Canino, G., Aguilar-Gaxiola, S., Febo, V., Magaña, C., Soto, J., Eisen, S.V.: Translation and cultural adaptation of a mental health outcome measure: the BASIS-R. Cult. Med. Psychiatry **31**(1), 25–49 (2007)

Eisen, S.V., Dill, D.L., Grob, M.C.: Reliability and validity of a brief patient-reported instrument for psychiatric outcome evaluation. Hosp. Community Psychiatry **45**, 242–247 (1994)

Eisen, S.V., Normand, S.L., Belanger, A.J., Spiro, A., Esch, D.: The revised Behavior and Symptom Identification Scale (BASIS-R). Med. Care **42**(12), 1230–1241 (2004)

Eisen, S.V., Gerena, M., Ranganathan, G., Esch, D., Idiculla, T.: Reliability and validity of the BASIS-24 mental health survey for whites, African-Americans, and Latinos. J. Behav. Health Ser. R. **33**(3), 304–323 (2006)

Eisen, S.V., Seal, P., Glickman, M.E., Cortés, D.E., Gerena, M.G., Aguilar-Gaxiola, S., Febo, V., Soto, J., Magaña, C., Canino, G.: Psychometric properties of the Spanish BASIS-24 mental health survey. J. Behav. Health Ser. R. (2009). doi:10.1007/s11414-009-9170-6

Fox, J.P., Glas, C.A.W.: Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika **66**, 269–286 (2001)

Gardner, W., Kelleher, K., Pajer, K.: Multidimensional adaptive testing for mental health problems in primary care. Med. Care **40**, 812–823 (2002)

Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/hierarchical Models. Cambridge University Press, New York (2007)

Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**, 457–511 (1992)
Gelman, A., Meng, X.L., Stern, H.S.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**, 733–807 (1996)
Geltman, D., Chang, G.: Hallucinations in Latino psychiatric outpatients: a preliminary investigation. Gen. Hosp. Psychiatry **26**(2), 153–157 (2004)
Glas, C.A.W.: Differential item functioning depending on general covariates. In: Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (eds.) Essays on Item Response Theory, pp. 131–148. Springer, New York (2001)
Glas, C.A.W., Meijer, R.: A Bayesian approach to person fit analysis in item response theory models. Appl. Psychol. Meas. **27**(3), 217–233 (2003)
Guarnaccia, P.J., Guevara, L.M., González, G., Canino, G., Bird, H.R.: Cross cultural aspects of psychotic symptoms in Puerto Rico. Res. Comm. Ment. Health **7**, 99–110 (1992)
Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models. Chapman and Hall, New York (1990)
Hoijtink, H.: Conditional independence and differential item functioning in the two-parameter logistic model. In: Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (eds.) Essays in Item Response Theory, pp. 109–129. Springer-Verlag, New York (2001)
Holland, P.W., Thayer, D.T.: Differential item functioning and the Mantel-Haenszel procedure. In: Wainer H., Braun, H.I. (eds.) Test Validity, pp. 129–145. Erlbaum, Hillsdale, NJ (1988)
Janssen, R., Tuerlinckx, F., Meulders, M., De Boeck, P.: A hierarchical IRT model for criterion-referenced measurement. J. Educ. Behav. Stat. **25**, 285–306 (2000)
Johnson, M.S., Sinharay, S.: Calibration of polytomous item families using Bayesian hierarchical modeling. Appl. Psychol. Meas. **29**, 369–400 (2005)
Junker, B.W.: Conditional association, essential independence and monotone unidimensional item response models. Ann. Stat. **3**, 1359–1378 (1993)
Kang, T., Cohen, A.S.: IRT model selection methods for dichotomous items. Appl. Psychol. Meas. **31**, 331–358 (2007)
Kok, F.G., Mellenbergh, G.J., van der Flier, H.: Detecting experimentally induced item bias using the iterative logit method. J. Educ. Meas. **22**, 295–303 (1985)
May, H.: A multilevel Bayesian item response theory method for scaling. J. Educ. Behav. Stat. **31**, 63–79 (2006)
Muraki, E.: A generalized partial credit model: application of an EM algorithm. Appl. Psychol. Meas. **16**, 159–176 (1992)
Pagano, I.S., Gotay, C.C.: Ethnic differential item functioning in the assessment of quality of life in cancer patients. Health Qual. Life Outcomes (2005). doi:10.1186/1477-7525-3-60
Patz, R.J., Junker, B.W.: A straightforward approach to Markov chain Monte Carlo methods for item response models. J. Educ. Behav. Stat. **24**, 146–178 (1999a)
Patz, R.J., Junker, B.W.: Applications and extensions of MCMC in IRT: multiple types, missing data, and rated responses. J. Educ. Behav. Stat. **24**, 342–366 (1999b)
Perkins, A.J., Stump, T.E., Monahan, P.O., McHorney, C.A.: Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. Qual. Life Res. **15**(3), 331–348 (2006)
R Development Core Team: R: A language and environment for statistical computing. (R Foundation for Statistical Computing), Vienna, Austria. http://www.R-project.org (2008)
Rosenthal, J.A.: Qualitative descriptors of strength of association and effect size. J. Soc. Service Res. **21**(4), 37–59 (1996)
Samejima, F.: Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monograph, No. 17 (1969)
Shealy, R., Stout, W.: A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika **58**, 159–194 (1993)
Sinharay, S.: Assessing fit of unidimensional item response theory models using a Bayesian approach. J. Educ. Meas. **42**(4), 375–394 (2005)
Spiegelhalter, D.J., Thomas, A., Best, N.G., Lunn, D.: WinBUGS 1.4 User Manual (Computer Program). MRC Biostatistics Unit, Cambridge, UK (2003)
Swaminathan, H., Rogers, H.J.: Detecting differential item functioning using the logistic regression procedures. J. Educ. Meas. **27**, 361–370 (1990)
Teresi, J.A.: Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Med. Care **44**, 152–170 (2006)

Thissen, D., Steinberg, L., Wainer, H.: Detection of differential item functioning using the parameters of IRT models. In: Holland, P.W., Wainer, H. (eds.) Differential Item Functioning, pp. 67–113. Erlbaum, Hillsdale, NJ (1993)

Thomas, A., O'Hara, B.O., Ligges, U., Sturtz, S.: OpenBUGS software package. R News **6**, 12–17 (2006)

Vega, W.A., Sribney, W.M., Miskimen, T.M., Escobar, J.I., Aguilar-Gaxiola, S.: Putative psychotic symptoms in the Mexican American population: prevalence and co-occurrence with psychiatric disorders. J. Nerv. Mental Dis. **194**(7), 471–477 (2006)

Wainer, H., Bradlow, E.T., Wang, X.: Testlet Response Theory and its Applications, chapter 14, pp. 219–233. Cambridge University Press, New York (2007)

Yee, T.W.: VGAM family functions for categorical data. Technical report, Department of Statistics, University of Auckland, New Zealand (2006)

Yee, T.W., Wild, C.J.: Vector generalized additive models. J. R. Stat. Soc. B **58**, 481–493 (1996)

Zhang, J., Stout, W.: Conditional covariance structure for generalized compensatory multidimensional items. Psychometrika **64**, 129–152 (1999)