A Bayesian IRT Model for Comparative Item Performance Under Dual Administration Modes

Louis T. Mariano, Maria Orlando, and Bonnie Ghosh-Dastidar
RAND Corporation

Abstract

Ordinal scale response items are often used in quantifying a latent trait. The mode in which these items are administered may effect an item's characteristics, such as the item's location on the latent scale and the efficiency of the item in discriminating between different values of the latent trait. We present the Bayesian Differential Mode Effects Model (BDMEM), a Bayesian Item Response Theory (IRT) model for the detection and quantification of mode of administration effects at both the item and form level. To illustrate the BDMEM, we present an example of a mental health survey administered both by telephone and self-administered questionnaire. The BDMEM is compared to the popular approach of IRT differential item functioning (DIF) evaluation, and its advantages over DIF are highlighted.

Keywords: Mode effect; Item Response Theory; Differential Item Functioning; Bayesian hierarchical models

1. Introduction

Ordinal scale response items are often used in quantifying a latent trait. For example, in the area of mental health, to better understand a patient's depression status, the patient may be administered a survey of depression-related items, responding to each item on a Likert-type scale. The latent trait may be parameterized using an Item Response Theory (IRT; e.g., van der Linden and Hambleton, 1997) model to describe the probability of endorsing the individual response categories of each item, conditional on the latent trait.

Administrations of a set of such items, often referred to as an item "scale" or "form," may occur in different modes of administration. A survey might be administered over the phone, face-to-face, by mail, or via the internet. We investigate the identification and quantification of mode of administration effects within a scale; i.e. whether administrations of a scale are equivalent in different modes.

While classic correlation and reliability analyses can provide some information about scale equivalence (e.g., Negy and Snyder, 2000; Norris and Perilla, 1996), an alternative and highly instructive conceptualization of the non-equivalence of scales focuses on the presence of statistical item bias, or differential item functioning (DIF). In this context an item is said to exhibit DIF if two respondents, who are administered a scale via different administration modes and have equal levels of the latent trait being measured, do not have the same probability of endorsing each response category of that item. The practical result at the item level is that scores on an item exhibiting DIF are not equivalent across modes. 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.

Subjects responding to a form in different modes create a natural grouping of the respondents. Investigation of item bias has been traditionally set in the context of examining items across gender or ethnic groupings. Unlike such groupings that are a function of the responding subject and deterministic in nature, the groupings defined by the mode of administration may be independent of the responding subject and assignable. Assignability of mode of administration allows for data collection designs which include the cross-classification of responding subjects to mode (i.e., individual subjects responding to subsets of items in different modes) and, where repeated responses are pragmatic, individual subjects responding to all items in both modes (i.e., a crossover design).

Below, we focus on the case where assignment in one of two modes of administration are possible, utilizing a design where data is collected from each subject in both modes. We contrast the likelihood ratio approach to IRT DIF evaluation, originally designed to test for item bias for the case of deterministic group assignment, with the Bayesian Differential Mode Effect Model (BDMEM) which we have developed to explicitly parameterize the mode effect and

Louis T. Mariano is Associate Statistician, Maria Orlando is Behavioral Scientist, Bonnie Ghosh-Dastidar is Statistician, RAND Corportaion, 1200 S. Hayes St., Arlington, VA 22202. Research supported in part by NIMH Grant #R03 MH62613 to Ghosh-Dastidar.

better accommodate more complex designs available under mode assignment.

2. Methodology

We draw upon Item Response Theory (IRT) to model the item characteristics and the latent trait. Specifically, we estimate the item category response function (ICRF), which describes the probability of responding in a given category on a given item, conditional on the latent trait. For ordinal categorical responses, one such model is Samejima's Graded Response Model (GRM, 1969):

$$\ln \frac{P(X_{ij} \geq k|\theta_i)}{1 - P(X_{ij} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk}). \qquad (1)$$

Here, θ_i represents the underlying latent trait of interest, X_{ij} is the response of subject i to item j, which falls into one of K ordinal response categories $k \in \{0, \ldots, K-1\}$. The $\gamma_{jk}, k \in \{1, \ldots, K-1\}$ reflect the spacing of the item responses along the latent construct continuum (Hambleton and Swaminathan, 1985) and thus determine the location of item j's ICRFs on the latent scale. The discrimination parameter, α_j, governs the slope of the ICRF, reflecting the degree to which the item is related to the underlying construct being measured. Higher values of α_j imply a strong relationship to the underlying construct. The ICRF is thus defined as:

$$P(X_{ij} = k|\theta_i) = P(X_{ij} \geq k|\theta_i) - P(X_{ij} \geq k+1|\theta_i),$$

$$\text{where } P(X_{ij} \geq k|\theta_i) = \frac{\exp\{\alpha_j(\theta_i - \gamma_{jk})\}}{1 + \exp\{\alpha_j(\theta_i - \gamma_{jk})\}}$$

The item-category location parameters γ_{jk} are nondecreasing in k and determine where the ICRF's for adjacent categories intersect; i.e., they determine where $P(X_{ij} = k|\theta_i) = P(X_{ij} = k-1|\theta_i)$. Model constraints are needed for identifiability; typically the latent trait is assumed to follow a standard normal distribution, $\theta \sim N(0, 1)$. The usual IRT assumptions of item independence (conditional on the latent trait) and unidimensionality of the latent trait also apply.

Other IRT modes for ordinal categorical data exist, such as Muraki's (1992) Generalized Partial Credit Model, which is similar to the GRM except the left hand side of the model is set as an adjacent category logit model (Agresti, 1990) instead of the cumulative logit form of the GRM.

2.1 Mode Effects

In considering the presence of mode of administration effects, we may express the hypothesis of interest in terms of the ICRF. Expanding the notation above, let X_{ijm} represent the response of subject i to item j when administered the item in mode $m \in \{1, 2\}$. Then, we may consider, for an individual item j:

$$\text{Ho:} \quad \forall(\theta_i, k), \quad P(X_{ijm} = k|\theta_i, m = 1) = P(X_{ijm} = k|\theta_i, m = 2),$$

versus the alternative of inequality.

The equality in the null hypothesis above may fail to hold due to a mode effect impacting item category location or discrimination. In the case of a location mode effect in category k, a level of bias, b_{jk}, would exist such that:

$$\forall\theta_i, \qquad P(X_{ijm} = k|\theta_i, m = 1) = P(X_{ijm} = k|\theta_i - b_{jk}, m = 2),$$

which is equivalent to increasing γ_{jk} by b_{jk} when $m = 2$. We concentrate on the identification of such location bias mode effects below.

The location bias b_{jk} may apply equally across the entire form or may differ across items; it may apply equally to all categories or a single category, at either the form or item level. These possible manifestations of location bias are specified in Table 1; multiple types may exist simultaneously.

2.2 Differential Item Functioning

Thissen, Steinberg, and Wainer (1993) provide an IRT likelihood ratio approach to DIF detection. We may apply this method to the GRM to investigate the presence of DIF within an individual item, comparing model fit between constrained and unconstrained versions of the GRM. The constrained version of the model simply ignores the mode of administration and applies the GRM as defined in Equation 1 above. In the the unconstrained version, the individual item is treated as two separate items, each with its own item-category parameter, γ_{jk1} and γ_{jk2} for each category k. The remaining items are assumed to function the same under both modes (i.e., each treated as a single item), and the GRM is fitted to the data, with the γ_{jk1} and γ_{jk2} determined by the information contained in the mode 1 and mode 2 responses respectively. The usual Chi-square Likelihood Ratio Test is then employed to determine whether the item parameters for item j are equal under the different modes; i.e., whether $\gamma_{jk1} = \gamma_{jk2}, \ \forall k$. Items are evaluated individually; a different version of the unconstrained model is constructed for each item which assumes no mode effect for any other item.

Table 1: Possible types of location bias caused by the mode of administration.

Form effects	$\forall(j,k),\ b_{jk} = b \neq 0$	bias is the same for all categories across all items
Category effects	$\exists k \ni \forall j,\ b_{jk} = b_k \neq 0$	bias is the same for a particular response category across all items
Item effects	$\exists j \ni \forall k,\ b_{jk} = b_j \neq 0$	bias is the same for all response categories within an individual item
Item-category effects	$\exists(j,k) \ni\ b_{jk} \neq 0$	bias is unique to an individual item category

Notice that the right-hand-side of Equation 1 includes a multiplicative term $\alpha_j\gamma_{jk}$. Testing for mode effects on one of these parameters remains meaningful only when the other is the same value across both modes. For example, if testing for location differences, the discrimination parameter α_j would be treated as the same across both modes in the unconstrained version.

The DIF method described above is a test for the existence of item or item-category location biases (Table 1). It does not test directly for form or category level effects (Table 1), nor does it quantify any of the possible effects it does detect.

As mentioned above, the GRM requires item independence given the latent trait. This could present an obstacle when an individual subject provides repeated measures, responding to an item in both modes of administration. In this case, the items represented by γ_{jk1} and γ_{jk2} would not be independent, and one of the responses would need to be discarded in order to fit the model. In the typical DIF application, with group membership defined by a characteristic of the responding subjects (such as race or gender), this is not a concern.

In the next section, we present a model that specifically parameterizes location bias at all four levels identified in Table 1 above, can examine all items for mode effects simultaneously, and can also accommodate responses from the same subject in both modes without violating the assumption of conditionally independent items.

2.3 The Bayesian Differential Mode Effects Model

We may quantify a location effect of the mode of administration by expanding the GRM of Equation (1) as follows:

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \psi_{jkm}), \quad (2)$$

where ψ_{jkm} is a general mode effects term capturing all four types of location bias described in Table 1:

$$\psi_{jkm} = \tau_m + \phi_{km} + \zeta_{jm} + \rho_{jkm}. \quad (3)$$

Following the parameter indices, τ_m represents form bias, ϕ_{km} represents category bias, ζ_{jm} represents item bias, and ρ_{jkm} represents item-category bias. Depending on the type of bias one wishes to investigate, these individual terms may be included or excluded from the model.

For example, a model for location bias at the form and category levels would set the ζ and ρ terms to zero; here, Equations (2) and (3) combine as:

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \tau_m - \phi_{km}). \quad (4)$$

Including a linear term to capture the additional variability attributable to the mode as in Equation (2) is related to the approach of Fischer's Linear Logistic Test Model (LLTM; 1973), later generalized to include an item-specific discrimination parameter by Patz and Junker (1999). The LLTM follows an adjacent category logit form, rather than the cumulative logit form of Equation (2).

We choose to cast Equations (2) and (3) in a Bayesian framework, that allows for consideration of a distribution of mode effects, instead of forcing any mode effects to be exactly the same across all possible administrations. Patz and Junker (1999) describe the general implementation of Bayesian IRT models. Equation 2 constitutes the likelihood portion of the model. With the population parameters of the latent variable fixed for model identifiability, prior distributions are set for the item and mode parameters (see Section 3.2 for an example). We may then explore the posterior distribution of the individual model parameters by sampling using Markov Chain Monte Carlo (MCMC; e.g., Gelman, Carlin, Stern and Rubin, 1995) techniques. We refer to this

Bayesian extension of the GRM that specifically parameterizes the potential location bias due to mode of administration as the Bayesian Differential Mode Effects Model (BDMEM). Note that the data level of the model may be set up in an adjacent category logit format, instead of the cumulative logit form of the GRM.

Additional model constraints are necessary to identify the BDMEM likelihood defined by Equations (2) and (3). Without loss of generality, we designate the mode represented by $m = 1$ as the reference mode. All mode effects for the reference mode are set to zero, $\psi_{jk1} \equiv 0$, and effects of the second (i.e., focal) mode ($m = 2$) are modeled relative to the reference mode.

If more than one type of location bias is included, a second mode constraint may be necessary. For example, in Equation (4), adding 1 to τ_2 and subtracting 1 from each of the $K-1$ ϕ_{k2} yields the exact same probability for each response category. Gaining identifiability may be accomplished by implementing a sum-to-zero constraint on the lower-order parameter. For Equation (4), the constraint would be placed on the category effect, as $\sum_{k=1}^{K-1} \phi_{k2} = 0$. Note that this constraint changes the meaning of ϕ_{km} to that of a categorical offset from the overall form-level mode effect attributable to all categories. The original definition of category effect may be recovered by summing the overall form effect and the constrained category offset.

The shortcomings of the DIF method examined in section 2.2 are overcome by the BDMEM. Form and category-level modes of administration bias may be directly evaluated. Instead of merely testing for such biases, the BDMEM quantifies them into posterior distributions of the mode effects. Since the BDMEM does not treat responses to the same item in different modes as responses to separate items, repeated measures may be used without violating the IRT assumption of the conditional independence of items.

3. Application

We examine survey response data to 10 depression related items, collected both via telephone interview and self-administered questionnaire (SAQ). The survey items available for analysis are a 10-item subset of the 23-item revised Center for Epidemiological Studies Depression Scale CES-D (Orlando, Sherbourne and Thissen, 2000), which was itself part of a broader interview. The 23-item version contains 13 items from the original CES-D (Radloff, 1977) and 10 new items. These 10 new items are the subset analyzed here. The items ask for the patients level of agreement with statements reflective of symptoms of depression, such as "I couldn't concentrate" or "I thought a lot about death." Responses are collected in four ordinal categories: "Rarely or none of the time," "A little of the time," "Occasionally," or "Most or all of the time."

The data were collected for a sub-study of the Partners-in-Care (PIC) Study, Depression Patient Outcomes Research Team-II (Wells, Sherbourne, Schoenbaum, et al., 2000), funded by the Agency for Healthcare Research and Quality. The PIC is a group-level, randomized controlled trial with the primary objective of determining whether quality-improvement interventions implemented in managed-care practices for depressed, primary-care patients improve quality of care, health and employment outcomes. The sub-study was conducted at the fourth follow-up wave of the PIC (18 months after baseline) to compare phone interview responses with those from SAQs.

A random subsample of 300 was selected from the 1,356 subjects enrolled in the PIC, and responses from $N = 246$ participating subjects are available. Half of this sample was randomly selected to receive the survey by phone interview first and then, 20 to 30 days later, respond to the form again in the SAQ format. The other half received the SAQ survey first, responding to the survey 20 to 30 days later via phone interview. Thus, we have responses to $J = 10$ items in each of two modes, phone and SAQ, with administrations occurring 20-30 days apart and first administration mode randomized. Four of the subjects have missing SAQ data; all subjects responded by phone. The period between administrations is small relative to the length of the study and we assume that the underlying latent depression of the subjects does not change over this period.

Below, we explore this data for the potential presence of mode of administration effects, between the phone and SAQ modes, first using the likelihood ratio DIF approach described in Section 2.2, then using the BDMEM introduced in Section 2.3.

3.1 DIF Analysis

As discussed in Section 2.2, the likelihood ratio DIF approach cannot accommodate repeated measures without violating the IRT conditional independence assumption. To explore the PIC sub-study data using this approach, we can only utilize a single response from each patient. We conduct the DIF analysis using the data from the first administration of the scale, then repeat the analysis using the second

administration data, looking to confirm the first administration results.

The constrained and unconstrained versions of the GRM were estimated using the Multilog program (Thissen, 1991). Multilog also carried out the likelihood ratio test (size = 0.05) for item or item-category mode effects, based on three degrees of freedom, as there were three additional item-category parameters in the unconstrained model.

Table 2 contains the results of the DIF tests for both administration samples. For the item-level tests of the 3 location parameters, results from the 2 administration samples agreed on 5 of the 10 items; items 1 and 10 did not exhibit item-level mode DIF in either sample, and items 2, 3, and 5 showed significant item DIF in both samples. Items 7-9 showed significant item DIF only in the first administration sample, while items 4 and 6 showed significant item DIF only in the second administration sample.

These results suggest a lack of power for the DIF identification, which is to be expected given the small number of observations (n=123) in each group, and emphasizes the weakness of this approach in only being able to utilize half the available data for any given likelihood ratio test. The implications from this DIF analysis for the potential existence of form-level mode effects are mixed and inconclusive.

3.2 BDMEM Analysis

In implementing the BDMEM on the PIC dataset, we first investigate mode effects at the form level, then at the individual item level, and then combine those results into a final model for mode effects.

For each version of the BDMEM, the population distribution of the latent parameters describing patients' depression is fixed at $\theta_i \sim N(0,1)$ for identifiability. Subject to any necessary additional model constraints, we choose Normal prior distributions for the item and mode parameters that are diffuse relative to the distribution of θ. Specifically, $\forall j,k\ \gamma_{jk} \sim N(0,10)$ and $\ln(\alpha_j) \sim N(0,10)$, and, when included, each of τ_2, ϕ_{k2}, ζ_{j2}, and ρ_{jk2} are also assigned a $N(0,10)$ prior. We treat administration by phone as the reference category ($m = 1$), estimating the effects of SAQ administration relative to phone.

For all versions of the model, sampling from the posterior distribution of the parameters follows a Metropolis-Hastings within Gibbs algorithm (e.g., Gelman, et al., 1995) programmed in C++. Convergence was assessed using the method of Gelman and Rubin (e.g., Gelman, et al., 1995), and all versions of the BDMEM discussed below converged in a maximum of 2,000 iterations. MCMC samples were drawn from an additional 8,000 iterations after a burn-in.

3.2.1 Mode Effects Across Items

To explore form-level mode effects, we set $\psi_{jkm} = \tau_m$, so that Equations (2) and (3) reduce to:

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \tau_m), \quad (5)$$

with $\tau_1 \equiv 0$. MCMC estimates of τ_2 indicate a posterior median of -0.164 and a 95% equal-tailed credible interval of $(-0.232, -0.095)$, which does not contain zero, indicating the presence of a mode of administration effect on all items. The negative sign on this parameter indicates that patients will be more likely to endorse higher valued response categories (those indicating a higher level of depression) when responding by SAQ. Using the posterior median as the Bayes estimator, these results imply a shift on all three ICRF's for all 10 items by -0.164 when these items are self-administered.

We next investigate whether this detected form effect applies equally to all categories by adding a category effect $\psi_{jkm} = \tau_m + \phi_{km}$ as in Equation (4), with $\sum_{k=1}^{K-1} \phi_{k2} = 0$ for identifiability. In the presence of category effects, the form effect no longer remains significant, with a 95% interval estimate of $(-1.119, 0.037)$. However, all three categorical effects are non-zero, indicating mode effects for the individual categories, present across all items. Thus, we eliminate τ_m, refitting the model with only the unconstrained category effects in the likelihood:

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \phi_{km}). \quad (6)$$

Table 3 contains the estimates of the unconstrained category effects. All three interval estimates exclude zero, indicating the presence of category effects. Further, note that the sign on these effects for $k = 1$ and 2 are both negative, while for $k = 3$ the sign is positive. Under SAQ mode, the IRCFs for the first two categories intersect at a lower value on the depression θ scale; this is also true for the second and third categories, however, the third and fourth IRCFs intersect at a higher value under SAQ mode.

To illustrate these effects, Figure 1 plots the ICRF's for item 6, in both the phone and SAQ modes. Under phone administration, the green line, representing the third response category, is never the modal ICRF; when administered by phone, this

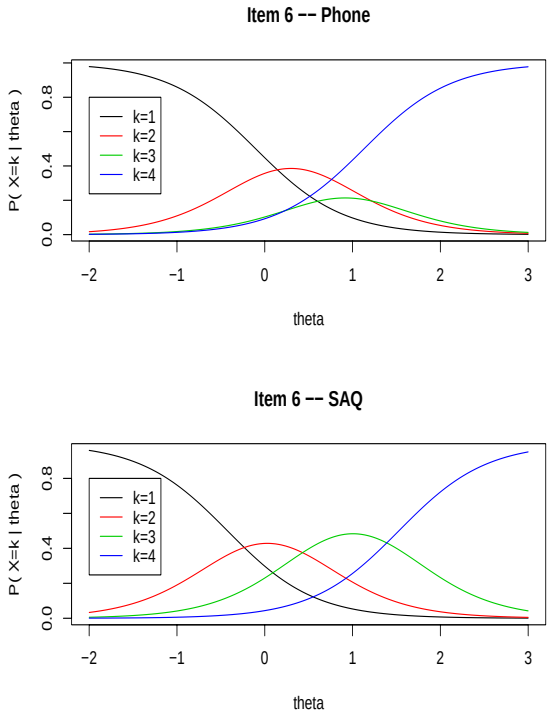
Table 2: Likelihood Ratio Test χ^2 statistics and p-values for mode of administration DIF for 10 items from the alternate CES-D scale, compared to the presence of item-category effects under the BDMEM.

	Item 1	Item 2	Item 3	Item 4	Item 5
1st DIF Administration	4.1 (0.251)	39.0 (0.000)	28.8 (0.000)	2.3 (0.681)	10.1 (0.039)
2nd DIF Administration	3.7 (0.296)	11.0 (0.012)	18.0 (0.001)	16.7 (0.002)	13.6 (0.009)
BDMEM item-category effects	Yes	Yes	Yes	Yes	Yes
	Item 6	Item 7	Item 8	Item 9	Item 10
1st DIF Administration	7.8 (0.099)	17.4 (0.002)	11.0 (0.027)	13.9 (0.008)	3.3 (0.509)
2nd DIF Administration	21.6 (0.000)	6.1 (0.192)	3.4 (0.493)	5.7 (0.223)	3.0 (0.558)
BDMEM item-category effects	Yes	Yes	No	Yes	No

Table 3: Posterior estimates of category mode of administration effects in the example CES-D subscale when administered via SAQ.

Parameter	Posterior Median	95% Credible Interval
ϕ_{12}	-0.321	(-0.416 , -0.233)
ϕ_{22}	-0.219	(-0.306 , -0.125)
ϕ_{32}	0.396	(0.280 , 0.525)

Figure 1: Estimated ICRF's for CES-D example item 6 under phone and SAQ modes of administration. Posterior median estimates of the item and mode parameters are used to generate the functions.



item is essentially functioning as a three-category response item, even though four categories are available. When the survey is administered via SAQ, all four categories are utilized.

3.2.2 Mode Effects Within Items

We next turn our attention to mode effects within an individual item j. As was the case at the form level, item effects attributable to all categories ζ_{jm} do not remain significant when specific item-category effects ρ_{jkm} are also included (results excluded for brevity). Thus, we focus on item-category effects present in the PIC data.

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \rho_{jkm}). \quad (7)$$

Table 2 compares these results to the DIF analysis. Of the 30 individual item-category mode effects, 13 are significant, as indicated by credible interval estimates that do not contain zero. These 13 span eight of the ten items. Comparing to the DIF results (Table 2), of the five items where the first and second administration DIF analyses agree, the BDMEM item-category results match four, with item 1 showing a significant mode effect only in the BDMEM analysis. This may be evidence of the loss of information in the DIF analysis in not being able to utilize the repeated measures. Of the five items where the first and second DIF analyses disagree, four of the five items show at least one significant item-category effect with the BDMEM, the exception being item 8. This false-positive might be attributible to the multiple testing problem present in the DIF analysis, as 20 size 0.05 DIF tests were carried out.

3.2.3 Combined Category Effects

At both the form and item levels, the overall mode effect attributable to the entire item, τ_m and ζ_{jm}

respectively, cease to be significant in the presence of category specific mode effects. This strongly indicates that the mode effects present are contained at the category level. Next, we include both form-category and item-category in the BDMEM, so that the data level of the model is described by:

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j(\theta_i - \gamma_{jk} - \phi_{km} - \rho_{jkm}). \quad (8)$$

In this version of the model, posterior estimates indicate that all but two of the item-category effects ρ_{jkm} do not remain significant in the presence of the category effects across items ϕ_{km}. Form this we conclude that mode of administration effects existing in this dataset are predominantly at the category level (Equation 6). Of course, formal Bayesian model selection techniques are available.

4. Discussion

We have introduced the Bayesian Differential Mode Effects Model, a Bayesian extension of the Graded Response Model specifically parameterized to capture the effects of administering a scale of items in two different modes. The BDMEM overcomes deficiencies found in the likelihood ratio DIF approach, by quantifying mode effects both within and across items and accommodating repeated measures.

In the CES-D example, we found the unexpected result of category mode effects dominating any effects present at the item level. This is counter to the idea that the more sensitive or stigmatizing items may show a stronger mode effect due to the added anonymity that the SAQ provides. It may be that there simply was not enough power, with 250 respondents, to detect these differences at the item level, however, the posterior median estimates at the item level seem to indicate otherwise. For example, using posterior median estimates, the mode effect for the prompt "I slept too much" was larger than for the prompt "I thought a lot about death." Of course, since the data for this sub-study were taken at the fourth follow-up wave of data collection, it is possible that the respondents were desensitized to the form by that point in the PIC study; it is also possible that item sensitivity is only apparent in the upper response categories.

Herein we have focused on mode of administration effects impacting the location of the ICRF. An extension of the BDMEM to account for mode effects on item discrimination are not difficult to envision; one such extension would multiply the item discrimination parameter α_j by a multiplicative factor ξ_{jkm} to account for the mode discrimination effect; e.g.,

$$\ln \frac{P(X_{ijm} \geq k|\theta_i)}{1 - P(X_{ijm} \geq k|\theta_i)} = \alpha_j \xi_{jkm}(\theta_i - \gamma_{jk} - \psi_{jkm}), \quad (9)$$

where ξ_{jkm} is set equal to 1 for the reference mode and estimated for the focal mode. This mode discrimination factor could be parsed into factors accounting for mode discrimination effects at the form, and item levels, for example, additively on the log scale.

Finally, while we have focused upon a repeated measures example to contextualize the development of the BDMEM, the model is valid for any DIF application where subjects are cross-classified with modes of administration, such that the sample of subjects cannot be parsed into unique groupings using [item `x` mode] combinations. Even in the case of fixed non-assignable group membership, the BDMEM would be a valid model for more traditional DIF applications with the same linking assumptions that DIF employs, such as the estimation of group means or the establishment of anchor items known a-priori to not be subject to mode effects. Here, group membership (e.g. male versus female) would play the role of the mode. While the advantage of cross-classification is lost, the BDMEM would still offer a quantification of the mode effect, as well as direct investigation of effects that span across items and the simultaneous investigation for effects within multiple items.

References

Agresti, A. (1990), *Categorical Data Analysis*, John Wiley and Sons, New York.

Fischer, G.H. (1973), "The Linear Logistic Test Model as an Instrument in Educational Research," *Acta Psychologica*, **37**, 359-374.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995), *Bayesian Data Analysis*, Chapman & Hall, London.

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

Muraki, E. (1992), "A Generalized Partial Credit Model: Application of an EM Algorithm," *Applied Psychological Measurement*, **16**, 159-176.

Negy, C. and 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. and 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., and Thissen, D. (2000), "Summed-score linking using item response theory: application to depression measurement," *Psychological Assessment*, **12**, 354-359.

Patz, R.J. and Junker, B.W. (1999), "Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses," *Journal of Educational and Behavioral Statistics*, **24**, 342-366.

Radloff, L.S. (1977), "The CES-D scale: A self-report depression scale for research in the general population," *Applied Psychological Measurement*, **1**, 385-401.

Samejima, F. (1969), "Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores," *Psychometrika Monograph*, No. 17.

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

Thissen, D., Steinberg, L., and Wainer, H. (1993), "Detection of Differential Item Functioning Using the Parameters of Item Response Models," In *Differential Item Functioning*, eds. P.W. Holland and H. Wainer, 67-114, Lawrence Erlbaum Associates, Hillsdale, NJ.

van der Linden, W.J. and Hambleton, R.K. (1997), "Item Response Theory: Brief History, Common Models and Extensions," In *Handbook of Modern Item Response Theory*, eds. W.J. van der Linden and R.K. Hambleton, 1-28, Springer-Verlag, New York.

Wells, K.B., Sherbourne, C., Schoenbaum, M., et al. (2000), "Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial," *Journal of the American Medical Association*, **283**, 212-220.