A Novel Method for Expediting the Development of Patient-Reported Outcome Measures and an Evaluation Across Several Populations

Lili Garrard, Larry R. Price, [...], and Byron J. Gajewski

Abstract

Item response theory (IRT) models provide an appropriate alternative to the classical ordinal confirmatory factor analysis (CFA) during the development of patient-reported outcome measures (PROMs). Current literature has identified the assessment of IRT model fit as both challenging and underdeveloped. This study evaluates the performance of Ordinal Bayesian Instrument Development (OBID), a Bayesian IRT model with a probit link function approach, through applications in two breast cancer-related instrument development studies. The primary focus is to investigate an appropriate method for comparing Bayesian IRT models in PROMs development. An exact Bayesian leave-one-out cross-validation (LOO-CV) approach is implemented to assess prior selection for the item discrimination parameter in the IRT model and subject content experts' bias (in a statistical sense and not to be confused with psychometric bias as in differential item functioning) toward the estimation of item-to-domain correlations. Results support the utilization of content subject experts' information in establishing evidence for construct validity when sample size is small. However, the incorporation of subject experts' content information in the OBID approach can be sensitive to the level of expertise of the recruited experts. More stringent efforts need to be invested in the appropriate selection of subject experts to efficiently use the OBID approach and reduce potential bias during PROMs development.

Keywords: OBID, Bayesian leave-one-out cross-validation, Bayesian IRT, Bayesian model comparison, patient-reported outcome measures, PROMs

Researchers often build a few candidate models and seek to select the most useful one for a given problem. The process of model comparison and selection requires rigorous model checking or assessment that is an integral part of any statistical analysis. In the development of psychometric instruments, apart from reliability, establishing evidence of validity is essential to ensuring an instrument's psychometric integrity. Developing an evidence-based argument that scores are accurate for their intended use requires acquiring data specific to content, construct, and predictive aspects (Nunnally & Bernstein, 1994). Historically, validity has been presented as three distinct but related components—content, criterion, and construct. Today validity is viewed as a unitary concept (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) where propositions for test score interpretation and use are supported by evidence unique to the measurement goal. Although developing a comprehensive picture of score validity often includes content and predictive components, construct validity receives the most attention from a statistical modeling perspective. This is because any score validity argument is impossible to make without evidence that the construct is relevant to the proposed interpretation and use of the scores.

Two approaches can be implemented to establish evidence of construct validity. When the participant sample size is adequately large, classical (i.e., frequentist) confirmatory factor analysis (CFA) is fairly reliable and easy to implement via statistical software such as Mplus or the free R package *lavaan* (Rosseel, 2012). Bayesian approach often becomes advantageous when classical CFA is challenged by small sample size (Gajewski, Price, Coffland, Boyle, & Bott, 2013; Garrard, Price, Bott, & Gajewski, 2015; Jiang et al., 2014), that may result in model convergence issues and unreliable parameter estimates.

An emerging topic in recent literature focuses on the development of patient-reported outcome measures (PROMs) or patient-reported outcome (PRO) instruments that often are designed as questionnaires with ordinal response options. PROMs have gained increasing public awareness in promoting patient-centered care, an important driving force behind the current U.S. health care. For instance, the pharmaceutical industry is required by the U.S. Department of Health and Human Services (DHHS) Food and Drug Administration (FDA) to submit evidence collected through PRO instruments in support of labeling claims. Detailed industry guidelines are provided by the FDA to assist pharmaceutical companies regarding the psychometric evaluation of any new or adapted PRO instruments (FDA, 2009).

Ordinal or binary (a special type of ordinal data) responses often are collected from PROMs that require a different modeling approach when compared with the classical CFA (e.g., normality assumption) for assessing an instrument's construct validity. Literature has shown that the categorical version of the classical CFA model with ordinal data is equivalent to a two-parameter item response theory (IRT) model with a probit link function, when all items on an instrument are ordinal (Johnson & Albert, 1999; Quinn, 2004). IRT parameter and person ability estimates are invariant (i.e., person ability estimates are not test dependent and item indices are not group-dependent; Hambleton, Swaminathan, & Rogers, 1991; Price, 2016). Importantly, the invariance property provides a way for uses of Ordinal Bayesian Instrument Development (OBID) to directly use or compare item and ability information acquired in one study to another. Assessing IRT model fit was considered as a challenging and underdeveloped area (Sinharay & Johnson, 2003; Sinharay, Johnson, & Stern, 2006). Recent advancement in the literature has shown increasing attention on using limited-information goodness-of-fit testing for IRT model fit (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Joe, 2005). This article extends the literature by focusing discussions around an alternative Bayesian IRT model comparison.

The fit of Bayesian models can be evaluated in several ways. One popular method is posterior predictive model checking (PPMC; Rubin, 1984), which is closely related to classical goodness-of-fit tests (Gelman, Meng, & Stern, 1996; Sinharay & Johnson, 2003). Other methods include graphical posterior predictive checks, assessing the posterior predictive *p* value, and/or the utilization of Bayes factors (Gelman, Hwang, & Vehtari, 2014). However, as pointed out by Gelman et al. (2014), when the objective is to compare models, the predictive model accuracy needs to be estimated. Cross-validation (CV) and information criteria measures are commonly used for Bayesian model comparison (Gelman et al., 2014; Vehtari & Lampinen, 2002; Vehtari & Ojanen, 2012). Information criteria are typically defined as deviance measures and represented by some variations of the log likelihood or log predictive density. Stone (1977) has showed the asymptotic equivalency between the two approaches such that information criteria can be viewed as approximations to various types of CV (Gelman et al., 2014).

The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002; Spiegelhalter, Best, Carlin, & van der Linde, 2014) remains a popular choice in the Bayesian literature despite criticisms and can be computed easily via the software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Viewed analogously to the Akaike information criterion (AIC; Akaike, 1973), DIC is considered as another pointwise measure for conditioning on the posterior mean, whereas AIC conditions on the maximum likelihood estimator. Watanabe (2010) recently proposed a more fully Bayesian approach, known as WAIC (widely applicable or Watanabe–Akaike information criterion). WAIC is considered more appealing than AIC and DIC as it not only conditions on the entire posterior distribution but also works well with hierarchical and mixture structure models (Gelman et al., 2014). Among other CV methods for evaluating out-of-sample prediction performance, Bayesian leave-one-out cross-validation (LOO-CV; Vehtari & Lampinen, 2002) has been shown to be asymptotically equivalent to WAIC (Watanabe, 2010) and more applicable to problems with small sample size (*n)*.

Although both WAIC and Bayesian LOO-CV exhibit appealing properties, they are applied less in practice as Bayesian CV approaches can become very computationally intensive due to Markov chain Monte Carlo (MCMC) simulation for all validation units. The computation burden might be tolerable for studies with smaller sample sizes. For large sample sizes, several approximation approaches have been proposed in the literature for Bayesian LOO-CV, such as importance sampling (IS; Gelfand, Dey, & Chang, 1992), expectation propagation and Laplace approximation (Vehtari, Tolvanen, Mononen, & Winther, 2014), Bayesian *K*-fold CV (Vehtari, Gelman, & Gabry, 2015), and a more recent Pareto smoothed importance sampling (PSIS) approach for regularizing importance weights (Vehtari & Gelman, 2015; Vehtari et al., 2015).

Within the context of latent variable modeling, excellent research recently has been conducted that approximates Bayesian LOO-CV for Gaussian latent variable models (Li, Qiu, Zhang, & Feng, 2014; Vehtari et al., 2014). However, common data collected from PROMs are ordinal in nature, which calls for an extension of the Gaussian model method to ordinal models (i.e., IRT models). Yet, there is a lack of Bayesian LOO-CV approximation with ordinal models in the current literature (A. Vehtari, personal communication, July 20, 2015). In addition, Bayesian model comparison should be evaluated from the perspective of prior selection for the IRT model parameters. The choice of prior distribution is relevant to posterior parameter inferences and model predictions when data are sparse (Gelman et al., 2014).

When developing PROMs for target populations with small sample sizes (e.g., in cases of rare disease), a novel method called OBID recently has been proposed to overcome the small sample size challenge, appropriately model participants' ordinal responses, and expedite the development of PROMs (Garrard et al., 2015). OBID is developed within a Bayesian two-parameter IRT with a probit link modeling framework. *Prior* distributions derived from content experts' data or prior studies (for establishing the instrument's content validity) are updated with participants' data to obtain a *posterior* distribution for IRT model parameters. Thus, OBID may alleviate the need for large sample sizes, especially for studies with target populations that are small to begin with. Reducing the number of participants will expedite the overall instrument development process and alleviate patients' burden.

The current work is motivated by the need to have an appropriate method for comparing Bayesian IRT models in PROMs development with the goal to expedite the development process when sample sizes become a concern (e.g., small and/or non-normally distributed data). The OBID approach is evaluated through real data applications, and the specific aims include (a) comparing the OBID models with both informative and flat priors using exact Bayesian LOO-CV, and (b) assessing subject content experts' bias through an exact CV information criterion (CVIC) measure. All real data used in the current study were collected for prior research purposes and provided to the authors in a de-identified fashion. Thus, this study was determined as non-human subject research by a Midwestern Academic Medical Center Internal Review Board (IRB).

Method

The main objective of this article is to evaluate further the OBID approach via Bayesian model comparison using real data applications. First, the OBID participant model and how an exact Bayesian LOO-CV can be applied to scenarios used in the current study will be briefly reviewed.

OBID Participant Model

OBID is an ordinal CFA-based approach under the Bayesian probabilistic framework. Continuing the notations from Garrard et al. (2015), a two-parameter IRT model with the probit link is expressed by

$$y_{ij} = c \text{ if } y_{ij}^{*} \in \left(T_{j(c-1)}, T_{jc}\right];\ i = 1, \ldots, N,\ j = 1, \ldots, P,\ c = 1, \ldots, C_j,$$

$$y_{ij}^{*} = \alpha_j + \lambda_j f_i + \varepsilon_{ij};\ f_i \sim N(0, 1),\ \varepsilon_{ij} \sim N(0, 1),\ i = 1, \ldots, N,\ j = 1, \ldots, P,$$

where y_{ij} represents the ith participant's response to the jth item; and C_j is the number of response options for the jth item. The ordinal response y_{ij} is related to a continuous latent variable y_{ij}^{*}, through a set of ordered cut-points T_{jc}, on y_{ij}^{*}. The two item-specific parameters are α_j, the negative difficulty parameter for the jth item, and λ_j, the discrimination parameter for item j. The latent ability variable f_i is constrained to follow a standard normal distribution with ε_{ij} being the measurement error. The model further can be interpreted such that the probability of a particular response option being endorsed depends on the probability that y_{ij}^{*} falls within an interval defined by the cut-points. Technical details on the ordinal IRT model were described by Albert (1992) and extended by Béguin and Glas (2001), Sahu (2002), and Culpepper (2015). Note that notations used here differ from the usual IRT notations in the psychometric literature.

Under the local independence or conditional item independence assumption (Price, 2016), the likelihood for the underlying continuous latent variable y_{ij}^{*} is

$$L(\boldsymbol{y}^{*} | \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{f}) = \prod_{i=1}^{N} \prod_{j=1}^{P} N\left(y_{ij}^{*} | \alpha_j + \lambda_j f_i,\ 1\right).$$

In the unidimensional (i.e., single-factor) OBID approach, the prior distribution of the item discrimination parameter λ_j is specified using content validity information from subject experts (i.e., item relevancy ratings; informative prior). Item relevancy commonly is rated by the experts using a 4-point relevancy scale (i.e., 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, 4 = *highly relevant*). Each expert's relevancy rating for each item is related to the same expert's latent item-to-domain correlation for the same item through either an equally spaced (i.e., [0.00, 0.25), [0.25, 0.50), [0.50, 0.75), and [0.75, 1.00), respectively) or an unequally spaced (i.e., [0.00, 0.10), [0.10, 0.30), [0.30, 0.50), and [0.50, 1.00), respectively) transformation on a latent correlation scale. Findings by Gajewski et al. (2012) suggest that for a panel of individuals with moderate level of expertise in the area of interest, the equally spaced transformation might be more appropriate. Interested readers are referred to Gajewski et al. (2012), Gajewski et al. (2013), Jiang et al. (2014), and Garrard et al. (2015) for additional background and details on the OBID approach.

Bayesian LOO-CV

Bayesian CV is a common method used to evaluate out-of-sample prediction performance and compare models. The idea behind CV is quite intuitive, and our description of the method intentionally is kept consistent with the work by Gelman et al. (2014) and Li et al. (2014). First, the full data set repeatedly can be partitioned into holdout data $\boldsymbol{y}_i$ and training data $\boldsymbol{y}_{-i}$. Because the focus is on LOO-CV, the holdout data set in our application will simply be a single participant's responses to all items on an instrument. Second, the model is fitted to the training data $\boldsymbol{y}_{-i}$, yielding the posterior distribution $P_{\text{post}(-i)}(\boldsymbol{\theta}, \boldsymbol{f} | \boldsymbol{y}_{-i})$ of the model parameters $\boldsymbol{\theta}$ and the latent variable $\boldsymbol{f}$, all denoted in the general notation format. Third, the posterior predictive density of the holdout data $\boldsymbol{y}_i$, conditioning on the training data, can be computed by specifying an evaluation function $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)$ that measures certain goodness of fit of the prediction to the actual holdout observation $\boldsymbol{y}_i$.

Following the work by Li et al. (2014), the CV posterior predictive evaluation is defined as the expectation of the evaluation function $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)$ with respect to the posterior distribution of the parameters, conditioning on the training data that can be expressed by

$$E_{\text{post}(-i)}\{a(\boldsymbol{y}_i, \theta, \boldsymbol{f}_i)\} = \int a(\boldsymbol{y}_i, \theta, \boldsymbol{f}_i) P_{\text{post}(-i)}(\theta, \boldsymbol{f} | \boldsymbol{y}_{-i}) \mathrm{d}\theta\, \mathrm{d}\boldsymbol{f}.$$

Suppose the evaluation function is taken as the value of the predictive density function at the actual holdout observation $\boldsymbol{y}_i$—that is, $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i) = P_{\text{pred}}(\boldsymbol{y}_i | \boldsymbol{\theta}, \boldsymbol{f}_i)$—the CV posterior predictive evaluation (Equation 4) becomes the CV posterior predictive density that can be approximated by averaging predictive densities at the actual holdout observation $\boldsymbol{y}_i$, across all MCMC draws from $P_{\text{post}(-i)}(\boldsymbol{\theta}, \boldsymbol{f} | \boldsymbol{y}_{-i})$. The CV posterior predictive density is expressed by

$$P_{\text{pred}}(\boldsymbol{y}_i | \boldsymbol{y}_{-i}) = \int P_{\text{pred}}(\boldsymbol{y}_i | \theta, \boldsymbol{f}_i) P_{\text{post}(-i)}(\theta, \boldsymbol{f} | \boldsymbol{y}_{-i}) \mathrm{d}\theta\, \mathrm{d}\boldsymbol{f},$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} P_{\text{pred}}^{s}\left(\boldsymbol{y}_i | \boldsymbol{\theta}^{s}, \boldsymbol{f}_i^{s}\right).$$

The participant model (Equations 1 and 2) in the current study is a single-factor two-parameter IRT model, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\lambda})$. For each sth MCMC posterior draw, estimates of the negative item difficulty parameter $\boldsymbol{\alpha}_j^{s}$, the item discrimination parameter $\boldsymbol{\lambda}_j^{s}$, and cut-points $\boldsymbol{T}_{jc}^{s}$ on $\boldsymbol{y}_{ij}^{*}$, can be obtained for each item on the instrument. In the current model, the latent variable $\boldsymbol{f}_i^{s}$ requires no updating, as MCMC draws come from the prior distribution. Then the predictive density at the actual holdout observation $\boldsymbol{y}_i$ at each MCMC iteration—that is, $\boldsymbol{P}_{\text{pred}}^{s}(\boldsymbol{y}_i | \boldsymbol{\alpha}^{s}, \boldsymbol{\lambda}^{s})$—can be computed as the multivariate normal distribution function evaluated on intervals defined by the cut-points of each item that is expressed by

$$P^{s}_{\text{pred}}\left(\boldsymbol{y}_i|\boldsymbol{\alpha}^{s},\boldsymbol{\lambda}^{s}\right)=\int_{T^{s}_{c-1}}^{T^{s}_{c}}\text{MVN}\left(\boldsymbol{y}^{*}_{i}|\boldsymbol{\alpha}^{s}+\boldsymbol{\lambda}^{s}f^{s}_{i},\boldsymbol{I}\right)\text{d}\boldsymbol{y}^{*}_{i}$$

Finally, the CVIC (Li et al., 2014) is computed by −2 times the sum of the log of the CV posterior predictive density, over all validation units. The model with the smaller CVIC value is preferred.

To demonstrate the computation of the CV posterior predictive density $P_{\text{pred}}(\boldsymbol{y}_i|\boldsymbol{y}_{-i})$ (Equations 5 and 6), a hypothetical three-item instrument with binary response options (i.e., 1/0 or correct/incorrect) will be used. Suppose the holdout data $\boldsymbol{y}_i$ represent the ith subject's responses to the three items, where $\boldsymbol{y}_i = (0, 1, 1)$. For items with binary response options, the single cut-point on the underlying continuous latent variable $\boldsymbol{y}^{*}_{ij}$ is zero. The binary response y_{ij} is related to the latent variable $\boldsymbol{y}^{*}_{ij}$ through the following function:

$$y_{ij}=\left\{\begin{matrix}0 & if\ y^{*}_{ij}\in(-\infty,0]\\0 & if\ y^{*}_{ij}\in(0,\infty]\end{matrix}\right\}.$$

Holdout data $\boldsymbol{y}_i$ can be used to determine the corresponding set of cut-points needed for each of the three items. For instance, the ith participant's actual response for the first item is 0; therefore, the set of cut-points used will be (−∞, 0]. At each sth MCMC iteration, the set of cut-points $(\boldsymbol{T}^{s}_{c-1},\boldsymbol{T}^{s}_{c}]$ can be specified, where $\boldsymbol{T}^{s}_{c-1}=(-\infty,0,0)$ and $\boldsymbol{T}^{s}_{c}=(0,\infty,\infty)$. The predictive density $\boldsymbol{P}^{s}_{\text{pred}}(\boldsymbol{y}_i|\boldsymbol{\alpha}^{s},\boldsymbol{\lambda}^{s})$ can be computed by evaluating the three-dimensional multivariate normal distribution function on intervals defined by these cut-points, using the R function *pmvnorm* (Genz et al., 2015; R Core Team, 2015). Finally, the CV posterior predictive density $P_{\text{pred}}(\boldsymbol{y}_i|\boldsymbol{y}_{-i})$ at the actual holdout observation $\boldsymbol{y}_i$ is approximated by averaging across all MCMC iterations. Additional computations are performed using the R package *MCMCpack* (Martin, Quinn, & Park, 2011; R Core Team, 2015) and WinBUGS (Lunn et al., 2000).

Real Data Applications

In this section, data collected from two breast cancer-related instrument development studies will be described and analyzed using the OBID approach. An exact Bayesian LOO-CV is applied to compare the choice of prior for the item discrimination parameter λ_j, and to assess subject experts' bias toward the item-to-domain correlation (or item relevancy), under both equally spaced and unequally spaced transformations.

Patient Assessment of Mammography Services (PAMS)–Short Form Satisfaction Survey

Routine utilization of mammography is the most widely recommended method for breast cancer screening and offers patients a chance of early detection that is critical for overall survival. However, potential factors, such as prior experiences and satisfaction with mammography, influence patients' decision on using mammography on a regular basis. The PAMS satisfaction survey was developed due to the lack of mammography-specific satisfaction assessments (Engelman et al., 2010; Engelman et al., 2016). The full PAMS survey consists of four factors with 20 items, and the PAMS-Short Form is a single factor with seven items. The seven items are designed to measure overall satisfaction. Other items on the full survey are added only when one needs to measure a specific domain (e.g., comfort). Items on the full survey are designed with scales ranging from two to six response categories. The seven short-form items can be rated on a 5-point Likert-type scale (i.e., 1 = *poor*, 2 = *fair*, 3 = *good*, 4 = *very good*, and 5 = *excellent*.)

PAMS experts and participants Six subject experts were consulted and instructed to rate the relevancy of each item (ranging from 1 = "not relevant" to 4 = "highly relevant") to the domain of interest. Recruited experts consist of individuals who have published or worked in some type of breast cancer research, including several physicians (Ndikum-Moffor et al., 2016). Participant data were collected from female patients to establish construct validity of the PAMS-Short Form instrument. Complete data (i.e., participants responded to all items) are used for the current study. Patients represented four ethnicity backgrounds: Hispanic (n = 36), Non-Hispanic White (n = 2,768), Black (n = 34), and American Indian (n = 287).

PAMS CV—Prior selection For this study, analyses focused on the Hispanic, Black, and American Indian populations. First, distribution of response options (potential range = 1 to 5) from the raw participant data were examined. Very few respondents selected poor to good response options; thus, a decision was made to collapse some response categories. Potential loss of information due to scale reduction is acknowledged; however, this decision should not affect the general trend in data. For Hispanic and Black data, the 5-point scale is reduced to a 3-point scale by collapsing poor, fair, and good response options; and poor to fair response options are collapsed for the American Indian data, turning the scale into a 4-point scale.

The OBID approach promotes the incorporation of content experts' information (when appropriate) for the item discrimination parameter λ_j. In the absence of subject experts or an appropriate prior reference data (Garrard et al., 2015), a flat prior—that is, $\lambda_j \sim N(0, 4)$—can be used to fit the model and obtain parameter estimates. Furthermore, an exact Bayesian LOO-CV is applied to compare the choice of using a flat prior versus an informative prior (under both transformations). CVIC values for the flat prior, the equally spaced transformation prior, and the unequally spaced transformation prior are 589.934, 503.681, and 482.870 for Hispanic; 598.064, 525.005, and 485.829 for Black; and 4,112.868, 4,042.250, and 4,054.876 for American Indian, respectively. Across all three populations, both informative priors are favored over the flat prior. Because models with smaller CVIC

values are preferred, results indicate that the unequally spaced transformation models are preferred for Hispanic and Black populations; whereas the equally spaced model appears to be slightly better than the unequally spaced model for the American Indian population.

PAMS CV—Expert bias It is beneficial to assess experts' bias toward the item-to-domain correlation (or item relevancy), especially for smaller sample sizes. Figure 1 displays the CVIC value for each selected number of experts K. CVIC is calculated by both randomly selecting one to five experts from the pool of six experts and artificially inflating the prior sample size to represent information from 12 experts. $K = 0$ implies the use of flat prior that is added to the plots for comparison purposes. As the number of experts increases, the majority of CVIC values under the unequally spaced transformation are smaller than that of the equally spaced transformation. The selected experts appear to be less biased for both Hispanic and Black populations. However, the same group of experts is slightly more biased for the American Indian population. The CVIC value sharply increases after five experts for the unequally spaced transformation, whereas the CVIC value continues to decrease for the equally spaced transformation. In addition, all equally spaced transformation plots indicate that six experts are adequate, which is consistent with the suggestion in the current literature (Polit & Beck, 2006).

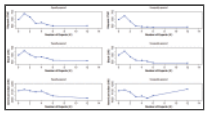

Figure 1.
PAMS expert bias comparison under both equally spaced (left panel) and unequally spaced (right panel) transformations.

Nutrition Literacy Assessment Instrument for Breast Cancer (NLit-BCa) Study

Motivated by a lack of nutrition literacy instrument for female breast cancer survivors, pilot work conducted by Gibbs et al. (2015) initiated the development of the NLit-BCa, an adapted version of the NLit (Gibbs & Chapman-Novakofski, 2013). The NLit-BCa consists of six individual domains with 75 items. A larger validity study is currently in process to evaluate further the NLit-BCa instrument (H. D. Gibbs, personal communication, August 25, 2015). Considering item revisions and/or deletions based on content experts' review, four domains with 39 items (i.e., 10 macronutrients [Macro] items, nine household food measurement [HFM] items, 10 food label and numeracy [FLN] items, and 10 consumer skills [CS] items) are deemed appropriate for analysis in this study. Items are designed with either three or four response options; and all participant responses are further classified as 0 = "incorrect" and 1 = "correct" based on an answer key provided by the instrument developers.

NLit-BCa experts and participants Four nutrition experts were consulted for the larger validation study and rated the relevancy for each of the 75 items. Recruited experts consist of individuals who have published expertise in cancer nutrition. Because the larger validation study is ongoing, participant data for this article will come from the pilot work. Data originally were collected from two groups of participants: weight loss intervention and non-intervention. Due to data sparsity concerns, complete data from 71 patients are used after combining both groups ($n = 25$ and 46 for the intervention and the non-intervention groups, respectively).

NLit-BCa CV—Prior selection A decision was made prior to analysis to exclude both Item 3 from the macronutrients domain (Macro03) and Item 2 from the FLN domain (FLN02) to avoid potential issues for LOO-CV analyses. Only one respondent answered Macro03 incorrectly, and everyone correctly answered FLN02. Thus, the total number of items was 37. The choice of flat prior versus an informative prior under both transformations was compared using exact Bayesian LOO-CV. CVIC values for the flat prior, equally spaced transformation prior, and the unequally spaced transformation prior are 504.101, 506.641, and 507.355 for Macro; 927.941, 947.594, and 941.578 for HFM; 633.835, 660.888, and 664.568 for FLN; and 716.069, 720.986, and 719.551 for CS, respectively. Across all four domains, the flat prior produces smaller CVIC values than both types of informative prior; however, the differences in CVIC values are much smaller for the CS domain.

NLit-BCa CV—Expert bias Results from the prior selection analysis seem to suggest that content experts are more biased toward item-to-domain correlations for all four domains. Figure 2 shows the CVIC value for each selected number of experts K. Similar to the PAMS study, the CVIC is calculated by both randomly selecting one to three experts from the pool of four experts and artificially inflating the prior sample size to represent information from eight experts. The use of flat prior again is indicated by $K = 0$. For the Macro domain, the CVIC value continues to decrease under the equally spaced transformation prior after four experts, where the opposite is observed with the unequally spaced transformation prior. No huge differences in CVIC values are observed among two to four experts for both HFM and FLN domains, under both transformations. For the CS domain, apart from the flat prior model, two experts produce the smallest CVIC value under the equally spaced transformation, whereas the smallest CVIC value occurs with three experts. Overall, recruited experts seem to be more biased toward relevancy ratings on the items, across all four domains.

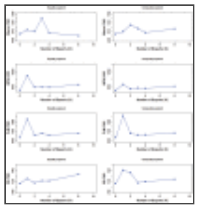

Figure 2.
NLit-BCa expert bias comparison under both equally spaced (left panel) and unequally spaced (right panel) transformations.

Discussion

The current study evaluates the performance of OBID through applications in two breast cancer-related instrument development studies. The primary focus is to investigate an exact Bayesian LOO-CV approach for comparing Bayesian IRT models in PROMs development. Six subject experts are consulted in the PAMS-short form study for four different patient populations. Among three populations investigated in this study, using an informative prior (i.e., incorporating experts' information), has shown to be superior to using a flat prior. One interesting observation arises from the original focus of the six content experts, as experts originally were recruited with the purpose of validating the PAMS instrument for American Indian women. Results from the PAMS study indicate that experts are less biased for both Hispanic and Black populations, which supports the appropriate utilization of experts' information to form a "general prior" as suggested by Garrard et al. (2015). Experts appear to be slightly more biased for the American Indian population despite their original focus. Although findings suggest that five experts would be sufficient, the use of six experts does not pose any substantial concerns for the purpose of instrument validation. Overall results indicate that incorporating information from the six selected subject experts is appropriate for the construct validity analysis in the PAMS study.

Findings from the NLit-BCa study present more complexity as the current study suggests the use of a flat prior as opposed to an informative prior. Among four domains examined, only the FLN domain CVIC results slightly support incorporating experts' information. The four selected experts appear to hold more biased opinions regarding the item-to-domain correlations for items in all domains. Although four experts were recruited, results have shown that even two to three experts would be sufficient. One thing worth noting is that the design of the NLit-BCa study differs from the PAMS study. The PAMS items are more subjective (i.e., eliciting satisfaction); whereas, the NLit-BCa items have a distinct correct answer. Nonetheless, despite the seemingly "opposite" results from the NLit-BCa study, the importance of appropriate prior selection and expert bias evaluation has been demonstrated for the OBID approach.

One limitation of the current study is associated with the selection of content experts, which remains an important yet challenging aspect in the development of psychometric instruments (Grant & Davis, 1997; Lynn, 1986). Apart from unidimensional instruments, subject experts often are asked to rate items from multiple domains. It usually is assumed that experts have expertise in all areas of interest. The current study assumes that content validity has been thoroughly assessed for both instruments. Thus, the focus is entirely on model selection during the construct validity phase of instrument development. Yet, based on current findings, subject experts' bias may hinder the efficient utilization of experts' information in the recently proposed OBID approach. Another limitation comes from the primary focus on using an exact Bayesian LOO-CV approach to compare different IRT models. As previously mentioned, several methods can be used to help assess and compare Bayesian models. The OBID approach certainly can be evaluated further via other established approaches in the literature. The third limitation can be viewed as a constraint associated with using the R package *MCMCpack*, as normal priors are required for IRT model parameters. Future work can consider other types of prior distributions.

An implication from the current study is the selection of an appropriate tuning parameter to ensure 20% to 50% acceptance rates during the MCMC procedure. Simulation results from Garrard et al. (2015) have showed an inverse relationship between the tuning parameter and the sample size. Although not discussed in the main text of the article, based on sample size information from 11 real data sets and four simulation data sets from Garrard et al., a power function is fitted for the tuning parameter t as a function of sample size n, that is, $t = 11.947n^{-.544}$ with $R^2 = .836$. This formula should be further refined as more data sets become available.

Additional future work may involve a more thorough evaluation of the equally spaced and unequally spaced transformations in other real applications and an approximation to the Bayesian LOO-CV for ordinal latent variable models. In addition, more skewed participant data structure and other prior distributions for the OBID subject experts' model need to be evaluated through simulation. The simulation study by Garrard et al. (2015) considers a more balanced participant data structure and that experts' item ratings follow a normal distribution. For instruments with more subjective response scales (e.g., satisfaction), participants tend to select more positive response options. Experts also potentially can disagree with each other regarding the relevancy of proposed items.

Acknowledgments

The authors thank both Dr. Kimberly Engelman for use of the Patient Assessment of Mammography Services (PAMS) data from Grants CCE-103763 and 5P20MD004805, and Dr. Heather Gibbs for use of the Nutrition Literacy Assessment Instrument for Breast Cancer (NLit-BCa) data from Grants

P30CA168524 and IRG-09-062-04.

Footnotes

Authors' Note: This article reflects the views of the authors and should not be construed to represent the U.S. Food and Drug Administration's (FDA) views or policies. Lili Garrard completed this work as a PhD student in the Department of Biostatistics at the University of Kansas Medical Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Conflicting Interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the National Institute of Nursing Research of the National Institutes of Health under Award R03NR013236.

Article information

Appl Psychol Meas. 2016 Oct; 40(7): 455–468.
Published online 2016 Jul 28. doi: 10.1177/0146621616652634
PMCID: PMC5029789
NIHMSID: NIHMS784195
PMID: 27667878

Lili Garrard,[1] Larry R. Price,[2] Marjorie J. Bott,[3] and Byron J. Gajewski[3,4]

[1]U.S. Food and Drug Administration, Silver Spring, MD, USA
[2]Texas State University, San Marcos, USA
[3]University of Kansas School of Nursing, Kansas City, USA
[4]University of Kansas Medical Center, Kansas City, USA
Lili Garrard, Division of Biometrics III, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, White Oak Building 21, Room 3672, Silver Spring, MD 20993, USA. Email: lili.garrard@fda.hhs.gov

References

1. Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov B. N., Csaki F., editors. (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281). Budapest, Hungary: Akademiai Kiado. [Google Scholar]
2. Albert J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational and Behavioral Statistics, 17, 251-269. [Google Scholar]
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association. [Google Scholar]
4. Béguin A. A., Glas C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika, 66, 541-561. [Google Scholar]
5. Cai L., Maydeu-Olivares A., Coffman D. L., Thissen D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. British Journal of Mathematical and Statistical Psychology, 59, 173-194. [PubMed] [Google Scholar]
6. Culpepper S. A. (2015). Revisiting the 4-parameter item response model: Bayesian estimation and application. Psychometrika. Advance online publication. doi:10.1007/s11336-015-9477-6 [PubMed] [CrossRef] [Google Scholar]
7. Engelman K. K., Daley C. M., Gajewski B. J., Ndikum-Moffor F., Faseru B., Braiuca S., . . . Greiner K. A. (2010). An assessment of American Indian women's mammography experiences. BMC Women's Health, 10, Article 34. doi:10.1186/1472-6874-10-34 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
8. Engelman K. K., Ndikum-Moffor F. M., Gajewski B. J., Yu Q., Nazir N., Daley C. M., Ellerbeck E. F. (2016). Reliability and validation of a patient assessment of mammography services (PAMS) satisfaction survey. Manuscript submitted for publication. [Google Scholar]
9. Gajewski B. J., Coffland V., Boyle D. K., Bott M., Price L. R., Leopold J., Dunton N. (2012). Assessing content validity through correlation and relevance tools: A Bayesian randomized equivalence experiment. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8, 81-96. doi:10.1027/1614-2241/a000040 [CrossRef] [Google Scholar]

10. Gajewski B. J., Price L. R., Coffland V., Boyle D. K., Bott M. J. (2013). Integrated analysis of content and construct validity of psychometric instruments. Quality & Quantity, 47, 57-78. doi:10.1007/s11135-011-9503-4 [CrossRef] [Google Scholar]
11. Garrard L., Price L. R., Bott M. J., Gajewski B. J. (2015). A novel method for expediting the development of patient-reported outcome measures and an evaluation of its performance via simulation. BMC Medical Research Methodology, 15(1), Article 77. doi:10.1186/s12874-015-0071-5 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
12. Gelfand A. E., Dey D. K., Chang H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In Bernardo J. M., Berger J. O., Dawid A. P., Smith A. F. M., editors. (Eds.), Bayesian statistics (4th ed., pp. 147-167). Oxford, UK: Oxford University Press. [Google Scholar]
13. Gelman A., Hwang J., Vehtari A. (2014). Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24, 997-1016. doi:10.1007/s11222-013-9416-2 [CrossRef] [Google Scholar]
14. Gelman A., Meng X. L., Stern H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6, 733-760. [Google Scholar]
15. Genz A., Bretz F., Miwa T., Mi X., Leisch F., Scheipl F., Hothorn T. (2015). mvtnorm: Multivariate normal and t distributions (R package version 1.0-3). Retrieved from http://CRAN.R-project.org/package=mvtnorm
16. Gibbs H. D., Chapman-Novakofski K. (2013). Establishing content validity for the Nutrition Literacy Assessment Instrument. Preventing Chronic Disease, 10, 120267. doi:10.5888/pcd10.120267 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
17. Gibbs H. D., Ellerbeck E. F., Befort C., Gajewski B., Kennett A. R., Yu Q., . . . Sullivan D. K. (2015). Measuring nutrition literacy in breast cancer patients: Development of a novel instrument. Journal of Cancer Education. Advance online publication. doi:10.1007/s13187-015-0851-y [PMC free article] [PubMed] [CrossRef] [Google Scholar]
18. Grant J. S., Davis L. L. (1997). Selection and use of content experts for instrument development. Research in Nursing & Health, 20, 269-274. [PubMed] [Google Scholar]
19. Hambleton R. K., Swaminathan H., Rogers H. J. (1991). Fundamentals of item response theory. Thousand Oaks, CA: Sage. [Google Scholar]
20. Jiang Y., Boyle D. K., Bott M. J., Wick J. A., Yu Q., Gajewski B. J. (2014). Expediting clinical and translational research via Bayesian instrument development. Applied Psychological Measurement, 38, 296-310. doi:10.1177/0146621613517165 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
21. Joe H., Maydeu-Olivares A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. Psychometrika, 75, 393-419. [Google Scholar]
22. Johnson V. E., Albert J. H. (1999). Ordinal data modeling. New York, NY: Springer Science & Business Media. [Google Scholar]
23. Li L., Qiu S., Zhang B., Feng C. X. (2014). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. Statistics and Computing. Advance online publication. doi:10.1007/s11222-015-9577-2 [CrossRef] [Google Scholar]
24. Lunn D. J., Thomas A., Best N., Spiegelhalter D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing, 10, 325-337. doi:10.1023/A:1008929526011 [CrossRef] [Google Scholar]
25. Lynn M. R. (1986). Determination and quantification of content validity. Nursing Research, 35, 382-385. [PubMed] [Google Scholar]
26. Martin A. D., Quinn K. M., Park J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software, 42(9), 1-21. [Google Scholar]
27. Maydeu-Olivares A., Joe H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. Journal of the American Statistical Association, 100, 1009-1020. [Google Scholar]
28. Ndikum-Moffor F. M., Braiuca S., Gajewski B. J., Daley C. M., Yu Q., Engelman K. K. (2016). Focus groups and content validity indexing utilization in the development of a patient assessment of mammography services instrument for American Indian women. Manuscript in preparation. [Google Scholar]
29. Nunnally I. H., Bernstein J. C. (1994). Psychometric theory. New York, NY: McGraw-Hill. [Google Scholar]
30. Polit D. F., Beck C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29, 489-497. [PubMed] [Google Scholar]
31. Price L. R. (2016). Psychometric methods: Theory into practice. New York, NY: Guilford Press. [Google Scholar]
32. Quinn K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. Political Analysis, 12, 338-353. doi:10.1093/pan/mph022 [CrossRef] [Google Scholar]
33. R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/
34. Rosseel Y. (2012). lavaan: An R Package for structural equation modeling. Journal of Statistical Software, 48(2), 1-36. [Google Scholar]
35. Rubin D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics, 12, 1151-1172. doi:10.1214/aos/1176346785 [CrossRef] [Google Scholar]
36. Sahu S. K. (2002). Bayesian estimation and model choice in item response models. Journal of Statistical Computation and Simulation, 72, 217-232. [Google Scholar]

37. Sinharay S., Johnson M. S. (2003). Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models (ETS RR-03-28). Princeton, NJ: Educational Testing Service. [Google Scholar]
38. Sinharay S., Johnson M. S., Stern H. S. (2006). Posterior predictive assessment of item response theory models. Applied Psychological Measurement, 30, 298-321. doi:10.1177/0146621605285517 [CrossRef] [Google Scholar]
39. Spiegelhalter D. J., Best N. G., Carlin B. P., van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B: Statistical Methodology, 64, 583-616. [Google Scholar]
40. Spiegelhalter D. J., Best N. G., Carlin B. P., van der Linde A. (2014). The deviance information criterion: 12 years on. Journal of the Royal Statistical Society, Series B: Statistical Methodology, 76, 485-493. doi:10.1111/rssb.12062 [CrossRef] [Google Scholar]
41. Stone M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, Series B: Statistical Methodology, 39, 44-47. [Google Scholar]
42. U.S. Department of Health and Human Services & Food and Drug Administration. (2009). Guidance for industry patient-reported outcome measures: Use in medical product development to support labeling claims. Retrieved from http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf [PMC free article] [PubMed]
43. Vehtari A., Gelman A. (2015). Pareto smoothed importance sampling. Retrieved from http://arxiv.org/pdf/1507.02646v2.pdf
44. Vehtari A., Gelman A., Gabry J. (2015). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. Retrieved from http://arxiv.org/pdf/1507.04544.pdf
45. Vehtari A., Lampinen J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation, 14, 2439-2468. doi:10.1162/08997660260293292 [PubMed] [CrossRef] [Google Scholar]
46. Vehtari A., Ojanen J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6, 142-228. [Google Scholar]
47. Vehtari A., Tolvanen V., Mononen T., Winther O. (2014). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. Retrieved from http://arxiv.org/pdf/1412.7461.pdf
48. Watanabe S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571-3594. [Google Scholar]