Published in final edited form as:

Appl Psychol Meas. 2014 June ; 38(4): 296–310. doi:10.1177/0146621613517165.

Expediting Clinical and Translational Research via Bayesian Instrument Development

Yu Jiang¹, Diane K. Boyle², Marjorie J. Bott³, Jo A. Wick¹, Qing Yu¹, and Byron J. Gajewski^{1,3,*}

¹Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, 66160

²University of Wyoming Fay W. Whitney School of Nursing, 1000 E. University Ave. Laramie, WY 82071

³University of Kansas School of Nursing, Kansas City, KS 66160

Abstract

Developing valid and reliable instruments is crucial but costly and time-consuming in health care research and evaluation. The Food and Drug Administration and the National Institutes of Health have set up guidelines for developing patient-reported outcome instruments. However, the guidelines are not applicable to cases of small sample sizes. Instead of using an exact estimation procedure to examine psychometric properties, our Bayesian Instrument Development (BID) method integrates expert data and participant data into a single seamless analysis. Using a novel set of priors, we use simulated data to compare BID to classical instrument development procedures and test the stability of BID. To display BID to non-statisticians, a graphical user interface based on R and WINBUGS is developed and demonstrated with data on a small sample of heart failure patients. Costs were saved by eliminating the need for unnecessary continuation of data collection for larger samples as required by the classical instrument development approach.

Keywords

BID; content validity; construct validity; patient reported outcomes; reliability

Introduction

Patient reported outcomes (PRO), which specifically refer to self-reports by patients, play an important role in health care research, system evaluation, and drug approval. Both the National Institutes of Health (NIH; Cella et al., 2010) and the Food and Drug Administration (FDA; Speight & Barendse, 2010) have set up guidelines for developing and accessing patient reported outcome measurement instruments. As implied in the extensive guidelines, classical psychometric instrument development is a lengthy process. Establishing a valid and reliable instrument requires necessary but time-consuming data gathering on content validity and on construct evidence. Typically, researchers begin the instrument development process

*Correspondence: Department of Biostatistics, University of Kansas Medical Center, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160 USA. Phone: 913-588-1603, Fax 913-588-0252, bgajewski@kumc.edu.

by first estimating content validity, which includes subject matter experts' evaluations of the extent to which items match the theoretical definition of the construct (Crocker & Algina, 1986). Estimating content validity is based on an interpretive trait-based argument, thereby providing a framework for linking the target domain to items that are developed and revised based on content expert opinions (Polit & Beck, 2006). Expert opinions only are used for making decisions about item contents (i.e., evidence of content validity), such as removing non-relevant items, fine tuning the wording of items, or adding new items. Next, through the collection of data from participants, the reliability and construct validity of items are tested by examining item interrelationships using factor analysis methods or an item response theory (IRT) model, which is an ordinal version of factor analysis from a modeling standpoint (Beck & Gable, 2001). In classical psychometric instrument development, the information provided by experts and by participants usually is analyzed separately. Information from the expert data (content analysis) is not used in a factor analysis (or IRT) of the participants' data (Messick, 1989).

Classical psychometric instrument development typically requires a large number of subjects (Westland, 2010; Westland, in press). Identifying sufficient numbers of subjects, particularly when there are limited resources and/or small populations, is not optimal for the timely development of needed instruments. For example, the development of an instrument for measuring outcomes for bipolar disorder took four years (Michalak & Murray, 2010). Therefore, a valid method that requires fewer subjects and less time is needed to expedite collecting data for research and translating findings into practice. The FDA and NIH do provide guidelines for factor analysis and IRT; however, they do not address situations for which targeted populations are small, which can arise when evaluating small racial or ethnic groups or patients suffering from rare disease. Lee (1981) developed a Bayesian approach for factor analysis and Lee and Song (2004) showed that it can be used for small sample sizes ($N=50$) while traditional factor analysis cannot. Gajewski et al. (2013) developed a unified psychometric model that potentially needs fewer subjects in estimating validity evidence for establishing new instruments. The Bayesian approach integrates the analyses of the content and construct validity (IACCV; Gajewski, Price, Coffland, Boyle, & Bott, 2013). In brief, the method first summarizes the state of knowledge of content validity using a *prior* probability distribution derived from expert judgment. The participant data are then used to update to a *posterior* distribution. IACCV was first used on an instrument measuring nursing home culture change (Gajewski, Price, Coffland, Boyle, & Bott, 2013). With this method, the investigators can transfer some of the response burden from the participants to an expert panel resulting in a faster, more efficient, and less costly instrument development process.

In this paper, we further advance IACCV methodology and report an easy-to-use software package that will implement our Bayesian instrument development method targeted for use by clinical researchers. Among the advancements, some are statistical and some are practical (e.g., software development). For example, in our previous work (Gajewski, Price, Coffland, Boyle, & Bott, 2013), a logit transformation was used for item-domain correlations, which is too restrictive because it produces only non-negative correlations. In the current study, we use Fisher's transformation (Wilks, 1962) because its range (-1 to 1) is a real representation of correlation. To reflect these new advancements in IACCV in the form of our new

software for clinical researchers, the current approach is called Bayesian instrument development (BID).

Evaluation and comparison of the Bayesian approach to likelihood methods has been studied by many researchers (Chaloner, 1987; Browne & Drapper, 2006). In general, a Bayes estimator has a smaller squared error but a larger bias compared to a maximum likelihood estimator in analysis of variance components (Chaloner, 1987). Comparing the Bayesian approach to traditional factor analysis was first investigated by Lee (1981) using simulation studies. The results showed that Bayes estimates using “reasonably good prior information” are better than flat priors, and both of the estimates have smaller root mean square error than the traditional maximum likelihood factor analysis approach (Lee & Shi, 2000). Samaniego and Reneau (1994) presented a landmark study which showed both theoretically and empirically that Bayesian estimates perform better than frequentists’ when the priors are close enough to the truth. However, when the priors are highly misinformed about parameters, Bayesian estimates may lose their superiority and should be processed with caution. We have extended the study into a comparison of BID and traditional factor analysis for various levels of expertise: some experts have a correct opinion and others have a wrong opinion. The effect of these “contaminated priors” on Bayesian factor analysis and its comparison to a classical approach is a main contribution of this paper.

Using simulated data, we test the BID approach by comparing it with classical instrument development in terms of performance stability and time consumption in development. The results will demonstrate that while the mean square error (MSE) of correlation estimation using classical instrument development does not change with respect to the number of experts, BID has lower error with even a single expert and further improves estimation efficiency as the number of experts increases. We will also demonstrate that the mean squared error for BID is smaller when compared to classical instrument development in the case when experts are biased by as much as 50% and the number of biased experts is small (e.g., 1 or 2 biased experts out of 6 total experts). To make BID user-friendly, we programmed a graphical user interface (GUI) version of BID using R and WinBUGS (R Development Core Team, 2012; Lunn, Thomas, Best, & Spiegelhalter, 2000). The GUI version of BID can guide users to retain widely accepted principles (e.g., reliability, factor structure, or item characteristics) of instrument development by simple point and click. The new BID software is demonstrated to clinical researchers who are non-statisticians and applied to a research project *Timeliness of Symptom Recognition, Interpretation, and Reporting in Heart Failure* (TSRIR) (K99NR012217).

BID model

Bayesian Instrument Development (BID) expands Integrated Analysis of Content and Construct Validity (IACCV) that was first developed by Gajewski et al. (2013). In the current paper, we will focus on a one factor (domain) BID model. For details of the general model, please see Gajewski et al. (2013). Because of its general use in practice and simple analytic form, we motivate the factor analytic model with a simple illustrative calculation of item-domain correlation.

Simple illustration: item to domain correlation

Suppose that the true distribution of item to domain correlation, after Fisher transformation, is normal; that is, $g(r) \sim N(g(\theta), \frac{1}{n})$, where r is the sample item to domain Pearson's correlation (e.g., standard calculation in psychometric software), n is the number of participants, and $g(\cdot)$ is a function, Fisher's transformation. Then, suppose the likelihood of correlation in a transformed scale can be written as $g(r)|g(\rho) \sim N(g(\rho), \frac{1}{n})$. The prior for experts has a normal distribution after Fisher transformation $g(\rho) \sim N(g(\rho_0), \frac{1}{n_0})$, where $g(\rho_0)$ is the transformed prior mean item to domain correlation and n_0 is the prior sample size. The posterior distribution of $g(\rho)$ is then $g(\rho)|g(r) \sim N(\frac{ng(r)+n_0g(\rho_0)}{n+n_0}, \frac{1}{n+n_0})$, where the mean is a linear combination of the transformed sample correlation and the prior mean, weighted by their respective sample sizes. The variance is simply the inverse of the sum of the observed and prior sample sizes.. After assuming a true but unknown fixed correlation, θ , we can calculate the mean squared error using the standard formula $MSE(E[g(\rho)]) = Var(E[g(\rho)]) + Bias^2$, where $E[g(\rho)]$ is the Bayes estimate for $g(\rho)$. Thus using standard expected value calculations we can see that

$$Bias = \frac{ng(\theta) + n_0g(\rho_0)}{n+n_0} - g(\theta) = \frac{n_0(g(\rho_0) - g(\theta))}{n+n_0}. \text{ Further, using standard variance calculation}$$

$$Var(E[g(\rho)]) = \frac{n^2}{(n+n_0)^2} \frac{1}{n} = \frac{n}{(n+n_0)^2}, \text{ and so the mean squared error is}$$

$$MSE(E[g(\rho)]) = Var(E[g(\rho)]) + Bias^2 = \frac{n}{(n+n_0)^2} + \frac{n_0^2}{(n+n_0)^2} (g(\rho_0) - g(\theta))^2. \text{ It is easy to show}$$

that using the MSE that BID is better than classic when $\frac{n}{(n+n_0)^2} + \frac{n_0^2}{(n+n_0)^2} (g(\rho_0) - g(\theta))^2 < \frac{1}{n}$, or

$(g(\rho_0) - g(\theta))^2 < (\frac{1}{n} - \frac{n}{(n+n_0)^2}) \frac{n_0^2}{n+n_0}$. Suppose the prior sample size is $n_0=30$ and data sample size is $n=77$. Using the inequality, BID is better than classical method when

$$(g(\rho_0) - g(\theta))^2 < \left(\frac{1}{n} - \frac{n}{(n+n_0)^2} \right) \frac{n_0^2}{n+n_0} = \left(\frac{1}{77} - \frac{77}{(77+30)^2} \right) \frac{30^2}{77+33} = 0.0527.$$

Missing from this illustrative example is a methodology for specifying the prior distribution as well as specification to a more general model like factor analysis. We explore these issues in the following sections and develop an estimation approach faster than Gajewski et al. (2013) for our simulations studies to calculate the mean squared error for a comparison of BID and the classical approach.

Model and computation for experts

For classical content validity, content experts are instructed to review all items and to "... indicate how relevant you perceive the item to be to the respective domain." For each item, the rating scale generally ranges from 1 indicating 'not relevant' to 4 signifying 'highly relevant'. A recent study showed that content validity can also be measured using correlation scales, which are equivalent to relevance scales (Gajewski et al., 2012). Therefore, in addition to relevancy, content validity can be further interpreted as the experts' opinion regarding item-to-domain correlation.

For the content expert data, suppose there are K experts ($k=1, 2, 3, \dots, K$) responding to p questions ($j=1, 2, 3, \dots, p$) and let x_{jk} represent the ordinal measure of the k^{th} expert's opinion for the j^{th} item's relevance. The correlation for the j^{th} item with its respective

domain as perceived by the k^{th} expert is denoted by ρ_{jk} and its assumed mapping from x_{jk} is as follows:

$$x_{jk} = \left\{ \begin{array}{ll} \text{'not relevant'} & \text{if } 0.00 \leq \rho_{jk} < 0.10 \\ \text{'somewhat relevant'} & \text{if } 0.10 \leq \rho_{jk} < 0.30 \\ \text{'quite relevant'} & \text{if } 0.30 \leq \rho_{jk} < 0.50 \\ \text{'highly relevant'} & \text{if } 0.50 \leq \rho_{jk} \leq 1.00 \end{array} \right\} \quad (1)$$

The correlation between the item and domain after pooling information from all experts is expressed as $\rho_j = \text{corr}(f, z_j)$, where f , the factor score of the domain, has a standard normal distribution with mean of zero and variance of one, and z_j is the standardized response for item j . The thresholds of ρ_j are argued under the assumption that the experts view correlations using Cohen's cut points (Cohen, 1988). Equation (1) allows us to properly model the uncertainty in these correlations that is induced by the interval censoring (e.g., < 0.10 , < 0.30 , etc.), similar to other methods used in studies on pressure ulcer staging (Gajewski, Hart, Bergquist, & Dunton, 2007), hearing test for audiology (Gajewski, Sedwick, & Antonelli, 2004; Gajewski, Nicholson, & Widen, 2009), and instrument development measuring nursing home quality of life (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Under some circumstances, if the prior data are from a panel knowledgeable in the area but not considered experts, an equal space transformation is found to be more appropriate and closer to actual correlations than the unequally spaced transformation (Gajewski et al., 2012; Pawlowicz et al., 2012). Equation (2) shows the equally spaced model:

$$x_{jk} = \left\{ \begin{array}{ll} \text{'not relevant'} & \text{if } 0.00 \leq \rho_{jk} < 0.25 \\ \text{'somewhat relevant'} & \text{if } 0.25 \leq \rho_{jk} < 0.50 \\ \text{'quite relevant'} & \text{if } 0.50 \leq \rho_{jk} < 0.75 \\ \text{'highly relevant'} & \text{if } 0.75 \leq \rho_{jk} \leq 1.00 \end{array} \right\} \quad (2)$$

Using Fisher's transformation, ρ_j is transformed to

$$g(\rho_j) = \frac{1}{2} \log \frac{1+\rho_j}{1-\rho_j} \quad (3)$$

A hierarchical model that combines experts and includes all items is:

$$g(\rho_{jk}) = g(\rho_j) + e_{jk}, \quad (4)$$

where e_{jk} is the random error assuming $e_{jk} \sim \text{i.i.d. } N(0, \sigma^2)$.

Assuming a lack of information on content validity prior to eliciting expert judgment, we use an essentially flat prior $\sigma^2 \sim \text{IG}(.00001, .00001)$ and $g(\rho_j) \sim N(0, 3)$. The shape and rate parameters for inverse gamma (IG) are both 0.00001. The prior specification allows us to estimate the posterior distribution of ρ_j using the expert's data via a Markov chain Monte Carlo (MCMC) procedure implemented in the software WinBUGS. These results then form

the prior distribution of ρ_j and are combined with the likelihood generated by participant data in the construct validity and reliability analysis. Note that this is a pool of expert's data for a simple, compromised prior, which is different from Samaniego & Reneau (1994), who analyze single expert at a time.

The previous model (Gajewski, Price, Coffland, Boyle, & Bott, 2013) for expert data utilized a logit transformation of correlation. However, because its range is 0 to 1 it can only be applied after reverse scoring items that are negatively correlated with the domain. For this reason, the logit transformation might be too restrictive. Using Fisher's transformation in the current model allows a real line representation of correlation from -1 to 1.

Based on our previous studies for nursing home culture change and nurse job satisfaction, we have found that the distribution of $g(\rho_j)$ can be approximated normal. The variance of $g(\rho_j)$ has an inverse relationship with the number of experts. That is, it is a simple representation of the sample size for the Bayesian prior. Specifically, the distribution of $g(\rho_j)$ can be written as

$$g(\rho_j) \sim N\left(g(\rho_{0j}), \frac{1}{n_{0j}}\right) \quad (5)$$

where $n_{0j} = 5 \times K$ and K is the total number of experts. Thus, the above equation is used in the simulation study. And it assumes each expert is worth five participants. The calculation of the variance and the justification of this estimation can be found in the Appendix. This approximation is not used in the actual BID software, but rather is used to study the one-factor analytic model's properties via simulation.

Model and computation for participants

For construct validity, we use a standardized factor analytic model to describe the item-to-domain correlations for participants' data (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Specifically, the one factor (domain) model is:

$$z_{ij} = \rho_j f_i + e_{ij} \quad (6)$$

Assuming y_{ij} is the i^{th} subject's response to the j^{th} item, it is standardized to z_{ij} , which allows us to focus on estimating item-to-domain correlations. The total number of participants is n and the total number of items is p ; f_i is the factor score for the domain (e.g., overall pain); and is typically assumed to follow a standard normal distribution, $N(0,1)$. The correlation between z_{ij} and f_i is represented by ρ_j ; thus, ρ_j^2 is the "reliability" of item j (Alonso, Laenen, Molenberghs, Geys, & Vangeneugden, 2010). The assumption $z_{ij} \sim N(0,1)$ implies $e_{ij} \sim N(0, 1 - \rho_j^2)$. Thus, the likelihood for z_{ij} is:

$$L(\mathbf{z}|\boldsymbol{\rho}, \mathbf{f}) = \prod_{i=1}^n \prod_{j=1}^p N\left(z_{ij}|\rho_j f_i, 1 - \rho_j^2\right) \quad (7)$$

As discussed previously, the distribution of $g(\rho_j)$ is approximately normal $g(\rho_j) \sim N(g(\rho_{j0}), \sigma_{g(\rho_{j0})}^2)$. To estimate the posterior distribution of the j^{th} component of $\boldsymbol{\rho}$, a flat prior $N(0, 3^2)$, representing high prior uncertainty in the construct validity of items can be used. An informative prior can also be utilized by obtaining the mean and variance of $g(\rho_{j0})$ from content analysis using method of moments. In our simulation study,

$g(\rho_j) \sim N\left(g(\rho_{j0}), \frac{1}{n_{j0}}\right)$ is used for the informative prior. Using this prior, the posterior distribution is:

$$\pi(\boldsymbol{\rho}, \mathbf{f} | n_0, \mathbf{z}) \propto \prod_{i=1}^n \prod_{j=1}^p \{N(Z_{ij} | \rho_j f_i, 1 - \rho_j^2)\} \times \prod_{i=1}^n N(f_i | 0, 1) \times \prod_{j=1}^p N(g(\rho_j) | g(\rho_{j0}), \frac{1}{n_{j0}}) / ((1 + \rho_j)(1 - \rho_j)). \quad (8)$$

The estimation of the posterior distribution of $\boldsymbol{\rho}$ can be obtained by MCMC methods performed in R and WinBUGS. However, this process is unnecessarily slow for simulation tests where only the posterior mode is needed. For example, when the number of items is 8 and the sample size is 100, it takes about 10 hours to compute 1,000 iterations using Equation (8) on a shared high speed workstation with dual Xeon 3.6 GHz processors and 4 GB of SDRAM.

Equation (8) provides the full posterior distribution that we use for clinical research, as it contains all the information for our inference. Although the posterior distribution does not have a closed form solution, it can be simplified because $\mathbf{z}_i$ follows a multivariate normal distribution, $\mathbf{z}_i \sim MVN(0, \mathbf{R})$. The variance of $\mathbf{z}_i$ is 1, the covariance between z_{ij} and z_{ik} is the product of ρ_i and ρ_j , and $\mathbf{R}$ is

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_1 \rho_2 & \cdots & \rho_1 \rho_p \\ \rho_2 \rho_1 & 1 & \cdots & \rho_2 \rho_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_p \rho_1 & \rho_p \rho_2 & \cdots & 1 \end{bmatrix}. \quad (9)$$

Then, the posterior distribution can be simplified as:

$$\pi(\boldsymbol{\rho}, \mathbf{f} | n_0, \mathbf{z}) \propto \prod_{i=1}^n MVN(\mathbf{z}_i | 0, \mathbf{R}) \times \prod_{j=1}^p N(g(\rho_j) | g(\rho_{j0}), \frac{1}{n_{j0}}) / ((1 + \rho_j)(1 - \rho_j)). \quad (10)$$

The specific estimation of ρ_j 's can be obtained by calculating the posterior mode of (10), which is the minimum of the following formula:

$$-\log(\pi(\boldsymbol{\rho} | n_0, \mathbf{z})) = -\sum_{i=1}^n MVN(\mathbf{z}_i | 0, \mathbf{R}) - \sum_{j=1}^p \log(N(g(\rho_j) | g(\rho_{j0}), \frac{1}{n_{j0}})) + \log(1 + \rho_j) + \log(1 - \rho_j) + C' \quad (11)$$

with constant, C. The estimation of ρ_j 's can be obtained by using the “nlminb” function in R, which is an optimization procedure using PORT routines (Bates & Sarkar, 2012). It has

been shown in the literature (Lee, 2007) that using multivariate-based equations, like Equation (11), for estimating the *full posterior distribution* is *slower* than using univariate-based equations, like Equation (8). However, we only need the *posterior mode* for our simulation study, so using Equation (11) is actually much faster and more convenient. For example, the processing time is *reduced* to about 5 minutes for the same analysis as described above using Equation (8).

Calculation of item reliability

In addition to validity, reliability is also an important concept in instrument development because it represents the precision of measurements. Cronbach's alpha has been widely used in the estimation of reliability. However, it is a lower bound estimator (Green & Yang, 2009), and therefore a conservative assessment of reliability. It is better to estimate composite reliability via a model based approach (Alonso, Laenen, Molenberghs, Geys, & Vangeneugden, 2010).

Under the one factor model, equation (6) also can be written in matrix form:

$$\mathbf{z}_i' = (z_{i1}, z_{i2}, \dots, z_{ip}) \text{ and } \boldsymbol{\rho}' = (\rho_1, \rho_2, \dots, \rho_p).$$

Because $e_{ij} \sim N(0, 1 - \rho_j^2)$, the residual covariance matrix for $\mathbf{e}_i$ can be written as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1-\rho_1^2 & 0 & \dots & 0 \\ 0 & 1-\rho_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1-\rho_p^2 \end{bmatrix}. \quad (12)$$

According to Alonso et al.(2010), the entire reliability is:

$$\mathbf{R}_\Lambda = 1 - \frac{|\boldsymbol{\Sigma}|}{|\mathbf{R}|}, \quad (13)$$

where $\mathbf{z}_i \sim MVN(0, \mathbf{R})$, and $\mathbf{e}_i \sim MVN(0, \boldsymbol{\Sigma})$. Entire reliability measures the reliability of the whole domain of items. Our extension using BID allows for efficient estimation of reliability for small populations.

Simulation study

We use simulated data to evaluate the performance of BID and compare it to the traditional factor analysis, a common approach for validating statistical methods. In BID, we assume all experts agree in interpreting the notion of correlation in their judgment response. Furthermore, experts' data are correlated with participants' data, indicating their judgments are the same or highly similar. To understand how violations of these assumptions may affect the analysis of data from expert opinion, we will test violations of the assumptions about performance of the experts. Using simulation, we will see more clearly the benefit-to-cost ratio of the number of experts needed for adequate evidence of content validity, and it will allow for exploration of the sensitivity of our assumptions regarding the experts' judgments.

According to our previous job satisfaction study, the average of the correlations ρ between individual item and domain is 0.5 (Gajewski, et. al., 2012); therefore, we will assume the true correlation is $\rho_{TRUE} = 0.50$ for all items. Under this assumption the scale score would explain 25% of the variation in the corresponding item. We obtain the Bayesian estimate—the posterior mode—using R instead of using the MCMC procedure in R and WINBUGS that estimates the full posterior distribution. The latter is used in our R-GUI BID software developed for clinical study researchers (see section 4). For comparison to traditional methods, classical factor analysis is done using the “Lavaan” package built in R (Rosseel, 2012). The numerical results of the “Lavaan” package are typically very close to that of the commercial package Mplus (Rosseel, 2010).

We test BID under two scenarios. First, we examine the stability of BID using different assumptions (unbiased experts $\rho_0=0.50$ or biased experts $\rho_0=0.75$), as the experts tend to overestimate the relevance of items in content validity (Gajewski, Price, Coffland, Boyle, & Bott, 2013). Specifically, we assume the prior distributions are $g(\rho_j) \sim N(g(0.5), n_{0j}^{-1})$ or $g(\rho_j) \sim N(g(0.75), n_{0j}^{-1})$. We then test BID performance when prior distributions (experts) are a mixture of unbiased and biased information.

Bias or unbiased priors

In order to test the performance of BID under biased prior expert information, we use a four factor simulation design. The first simulation factor is the presence or absence of expert bias—either (1) the judgment of correlation between experts provides estimates of the square root of item reliability ($\rho_0=0.50$, unbiased for all items), or (2) the experts are biased ($\rho_0=0.75$, for all items). We also examine the impact of bias over a range of number of items (8, 16, or 24), number of subjects (50, 100, or 200) and number of experts (1, 2, 6, or 16), which are typical instrument development studies with small sample sizes (Polit & Beck, 2006).

We use R to generate participant data based on the assumption that the standardized item response z_i is multivariate normal with mean vector zero and variance of identity. Each factor combination simulation is repeated for 1,000 times. Let $\hat{\rho}_j(s)$ be the posterior mode for

the s^{th} simulation in a total of 1,000 iterations. Then $MSE_j = \sum_{s=1}^{1000} (\hat{\rho}_j(s) - \rho_j)^2 / 1000$ and $\overline{MSE} = p^{-1} \sum_j MSE_j$; $Bias_j^2 = \left(\sum_{j=1}^{1000} (\hat{\rho}_j(s) / 1000 - \rho_{TRUE}) \right)^2$; and $\overline{Bias^2} = p^{-1} \sum_j Bias_j^2$. In our study, the $\overline{MSE}$ and $\overline{Bias^2}$ are used for comparison via graphs.

Figure 1 shows the MSE correlation estimates for unbiased and biased experts with participant sample sizes of 50, 100, or 200 when the number of items is 8. As traditional factor analysis does not use content validity information, it is not surprising that the MSE is unaffected by prior information. When the priors are unbiased, the MSE obtained by BID is always smaller than classical factor analysis when the sample size is 50 or 100, indicating BID is more efficient. When the sample size is large ($n=200$), the MSE by classical factor

analysis is slightly smaller than that of BID only when the number of experts is one. As the number of experts increases, the gains in efficiency by BID also increase. However, when the priors are biased, the relative efficiency of BID compared to classical factor analysis is a function of the number of experts. When the number of experts is small (e.g., one or two), BID has lower MSE than traditional methods. As the number of experts increases, any gains in efficiency recede and eventually traditional methods become superior. When sample size is smaller (e.g., $n=50$), efficiencies gained by BID are most obvious. The differences in MSE when using BID versus classical factor analysis follow a similar pattern when the number of items is 16 or 32. The corresponding figures are shown in the supplemental sections.

Overall, our first simulation study shows the impact of biased experts on the efficiency of BID relative to traditional methods is a function of the number of biased experts. When the expert data are unbiased, the MSE is always smaller when using BID rather than classical factor analysis for small sample studies. However, a stubbornly biased subjective prior is quite harmful. Our findings are consistent with those discussed in other Bayesian literature investigating Bayesian and classical estimation (Samaniego and Reneau 1994, Garthwaite, Kadane, & O'Hagan, 2005).

Contaminated priors

A second simulation study assessed the impact of having a contaminated expert panel; that is, assuming the number of experts is six and we investigated the impact of having one or two with biased judgment. Here, we let c represent the number of biased experts and examine the effects of prior contamination for participant sample sizes of 50, 100, and 200; and numbers of items of 8, 16, and 24.

We investigated the theoretical distribution of $g(\boldsymbol{\rho})$ when the total number of experts is 6, and the number of biased experts ($\rho_0=0.75$) is $c=0, 1, 2$, or 6, respectively. When all experts are unbiased ($\rho_0=0.50$), the distribution of $g(\boldsymbol{\rho})$ is normal with mean $0.5\log(3)$ and variance $1/30$. The distribution of $g(\boldsymbol{\rho})$ shifts to the right when there is one biased expert in the panel. The upper tail becomes heavier when contaminated priors increase to two. When all the experts are biased, the distribution of $g(\boldsymbol{\rho})$ changes back to a normal distribution with mean $0.5\log(7)$, and variance $1/30$. The theoretical distributions of $g(\boldsymbol{\rho})$ and $\boldsymbol{\rho}$ are shown in supplementary figures S1A and S1B.

Similar to the first set of simulations, R was used to generate data and calculate the posterior mode for $\boldsymbol{\rho}$, and 1,000 iterations were carried out for each simulation scenario. The MSE shown in the graph is $\overline{MSE}$ from the 1,000 replications.

As shown in Figure 2A, compared to classical factor analysis, BID always has a lower MSE of correlation estimates when the number of biased experts is one or two for a panel with six experts. Fewer contaminated priors are always better, as the data shows the MSE with only one biased expert is lower than that of two biased experts. The results are very similar across participant sample sizes of 50, 100 and 200, but the gain in efficiency decreases as the number of participants increase. Our results show that when the number of items is 8, the MSE of BID with one contaminated priors is equivalent to the MSE from a classical CFA method ($MSE = 0.01$). However, the level of precision is achievable with half the number of

participants ($n = 100$ for CFA, $n = 50$ for BID). Figure 2B shows the bias² of correlation estimates. The square of bias is larger using BID when there are biased experts in the panel compared to the classical method. The results of bias² indicated that BID lowers MSE through decrease of variance. Overall, estimation using informative priors is more efficient, which confirms the novel finding of Samaniego and Reneau (1994).

Application to data from heart failure study and development of GUI-BID software

We programmed a graphical user interface (GUI) version of BID using “TclTk” and “fgui” package in R and WinBUGS (Grosjean, 2012; Thomas & Nan, 2009). The new user-friendly BID software was readily usable by clinical researchers who were non-statisticians. An example of the GUI-BID software window can be found in figure S-4. The clinical researchers can choose to use a flat prior or an informative prior from expert data by selecting the option “prior” to be true or false. Data loading for both expert and participant data is accomplished by using the BID software in a point and click environment. The users are also guided to add the factor structure for the instrument. Although we focused on a one factor model for this paper, the software applies to multi-factor models as well.

In a study of heart failure, we apply BID to develop an instrument with the purpose to assess symptoms in heart failure patients. Using the proposed instrument, *Timeliness of Symptom Recognition, Interpretation, and Reporting in Heart Failure (TSRIR)* (K99NR012217), the study is conducted by nursing researchers to examine the symptom experiences of patients with heart failure leading up to hospitalization for acute decompensated heart failure. The tested instrument has ten items, and the experts’ ratings are on a 1–4 scale, from “not relevant” to “highly relevant”.

First, the experts’ data are analyzed using the Bayesian approach with equal space transformation discussed in section 3. In the MCMC procedure for estimating ρ_j we used three chains and a burn-in of the first 2,000 draws of each chain. The next 10,000 iterations were used for inferences. We also perform BID on participant data using a flat prior; that is, using $g(\rho_j) \sim N(0, 3^2)$ in the BID model. The results are shown in Table S-1. The entire reliability using the flat prior specification is 0.943, and its standard deviation is 0.010. Using an informative prior, $g(\rho_j) \sim N(g(\rho_{j0}), \sigma_{g(\rho_{j0})}^2)$, with an equal spaced model, the total reliability is 0.938 with standard deviation equal to 0.007. According to NIH PROMIS cut points, the reliability is larger than 0.8 and should be considered substantial (Cella et al., 2010). Overall, in the heart failure study the dimensions (scales) were deemed reliable with a small sample size ($n=60$) using BID.

Discussion

Classical instrument development is a powerful tool in patient reported outcome measurements when one has easy access to a large number of participants and considerable resources. However, in practice it usually happens that either the numbers of participants or resources are limited. Using a Bayesian approach, BID provides an innovative method for instrument development. BID is particularly valuable for studies with limited participants,

such as studies involving minority groups or those with orphan or rare diseases. In the heart failure study, it is notable that it took over two years to obtain data from 60 patients. The ratio of the variance of reliability by flat prior to informative prior is $0.01^2/0.007^2=2.04$. Thus, 104% more data or 123 participants would be needed if using traditional approaches. Consequently, it took two years to gain this clinical data, and it would take at least two more years without the informative prior. Using the BID approach, the clinical researchers do not need to continue data collection to obtain a larger sample that would be required by the classical instrument development approach. Both cost and time savings are achieved. Even when large samples are available, limited resources still may present challenges for investigators. For example, the development of an instrument measuring smoking cessation practices in methadone treatment clinics took four years because funds were limited for both participant incentives and personnel. If using BID, the processes could speed up dramatically (Cupertino et al., in review). Overall, BID proposes a psychometrically comprehensive and statistically efficient method for instrument development.

Using indirect measurement information is an important trend in statistics and related areas (Efron 2010). Samaniego and Reneau (1994) revealed that a good utilization of this indirect information, in the form of a prior distribution, can lead to a Bayesian estimator which is superior to a frequentist estimator in most situations. We extended their study into factor analysis for estimation of construct validity. In the current paper, we provided a full comparison of Bayesian approach and traditional factor analysis using simulation studies. In this paper, we investigated the performance of the Bayesian approach under different conditions: (1) the priors (experts) are all unbiased, (2) the priors (experts) are all biased (3) the priors are contaminated, with 1 or 2 biased experts among a total of 6 experts. Our results are consistent with what was in Lee (1981, Lee & Shi 2000), which shows that the Bayesian approach performs better than traditional factor analysis when using a flat or “good informative” prior. Our results are consistent with Samaniego and Reneau (1994)’s general finding that a Bayesian estimate is always superior to a frequentist estimate if the bias in the Bayesian prior is smaller than the sample standard deviation. In addition, we showed more clearly that the Bayesian approach is practically better than traditional methods, especially for small sample sizes as in our heart failure example.

Although the current study focuses on a one factor model, BID can easily be applied to multi-factor models. For the heart failure study data, we also tried different factor structures (i.e., using multiple factors and/or fewer items). The results were consistent with the expected results. In the BID model, we assume the experts’ opinion on items is independent and our prior is weighted over all the experts. The assumptions were used to simplify the model. However, in real circumstances the experts’ opinions on items are most likely correlated. In future studies, a correlated model may provide a better estimation for the content validity. A hierarchical model could be used to allow for individualized priors for each expert based on their prior beliefs. This would provide us with the ability to evaluate the effect of a single expert on the analysis and draw conclusions, using an approach similar as discussed by Samaniego & Reneau (1994). In BID, we also assume the participants’ data are continuous; however, many clinical questionnaires are in ordinal or binary form. Therefore, Ordinal Bayesian Instrument Development (OBID) using item response theory is a topic for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thanks to Drs. K Reeder and Carol Smith for use of their data from grant K99NR012217. This work was supported by a CTSA grant from NCRR and NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # UL1TR000001 (formerly #UL1RR033179). The data described is part of a clinical trial supported by the National Institute of Heart, Lung and Blood (NHLBI) Grant # HL 085397, C. Smith RN PhD, PI. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NCRR, or NCATS. The first author was partly supported by P20MD004805 Center for the American Indian Community Health (CAICH).

References

1. Alonso A, Laenen A, Molenberghs G, Geys H, Vangeneugden T. A Unified Approach to Multi-item Reliability. *Biometrics*. 2010; 66:1061–1068. [PubMed: 20070298]
2. Bates, D.; Sarkar, D. [accessed on April] R port. 2012. <http://netlib.bell-labs.com/netlib/port/>
3. Beck CT, Gable RK. Item response theory in affective instrument development: an illustration. *Journal of Nursing Measurement*. 2001; 9:5–22. Review. [PubMed: 11469142]
4. Broemeling, LD. Bayesian Biostatistics and Diagnostic Medicine. Boca Raton, FL: Chapman & Hall/CRC; 2007.
5. Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 1. 2006:473–550.
6. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays RD. on behalf of the PROMIS Cooperative Group. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*. 2010; 63:1179–94. [PubMed: 20685078]
7. Chaloner K. A Bayesian Approach to the Estimation of. Variance Components for the Unbalanced. One-Way Random Model, *Techn*. 1987; 29:322–337.
8. Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
9. Crocker, L.; Algina, J. Introduction to Classical and Modern Test Theory. Boston, MA: Harcourt, Brace, Jovanovich; 1986.
10. Cupertino, AP.; Hunt, JJ.; Gajewski, BJ.; Jiang, Y.; Marquis, J.; Friedmann, PD.; Engelman, KK.; Richter, KP. Substance Abuse Treatment, Prevention, and Policy. (in review)
11. Efron B. The future of indirect evidence. *Statistical Science*. 2010; 25:145–157. [PubMed: 21243111]
12. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. New York: Chapman and Hall; 2004.
13. Gajewski BJ, Sedwick JD, Antonelli PJ. A log-normal distribution model of the effect of bacteria and ear fenestration on hearing loss: A Bayesian approach. *Statistics in Medicine*. 2004; 23:493–508. [PubMed: 14748041]
14. Gajewski BJ, Hart S, Bergquist S, Dunton N. Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*. 2007; 26:4602–4618. [PubMed: 17393413]
15. Gajewski BJ, Nicholson N, Widen JE. Predicting hearing threshold in non-responsive subjects using a log-normal Bayesian linear model in the presence of left censored covariates. *Statistics in Biopharmaceutical Research*. 2009; 1:137–148.
16. Gajewski BJ, Price L, Coffland V, Boyle D, Bott MJ. Integrated analysis of content and construct validity of psychometric instruments. *Quality & Quantity*. 2013; 47(1):57–78.
17. Gajewski BJ, Coffland V, Boyle D, Bott MJ, Price L, Leopold J, Dunton N. 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. 2012; 8:81–96.
18. Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*. 2005; 100:680–701.
 19. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modeling. *Statistician*. 1994; 43:169–177.
 20. Green SB, Yang Y. Commentary on coefficient alpha: A cautionary tale. *Psychometrika*. 2009; 74:121–135.
 21. Grosjean, P. SciViews: A GUI API for R. UMONS; Mons, Belgium; 2012. URL <http://www.sciviews.org/SciViews-R>
 22. Lee SY. A Bayesian approach to confirmatory factor analysis. *Psychometrika*. 1981; 46:153–160.
 23. Lee, SY. *Structural Equation Modeling: A Bayesian Approach*. Wiley; New York; 2007.
 24. Lee SY, Shi JQ. Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annals of the Institute of Statistical Mathematics*. 2000; 52:722–736.
 25. Lee SY, Song XY. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*. 2004; 39:653–686. 2004.
 26. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS--a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*. 2000; 10:325–337.
 27. Michalak EE, Murray G. Development of the QoL. BD: a disorder-specific scale to assess quality of life in bipolar disorder. *Bipolar Disorder*. 2010; 12:727–40.
 28. Messick, S. Validity. In: Linn, RL., editor. *Educational Measurement*. 3. New York: American Council on Education and Macmillan; 1989. p. 13-103.
 29. Pawlowicz F, Gajewski BJ, Coffland V, Boyle D, Bott M, Dunton N. Application of Bayesian methodology within an equivalence content validity study. *Nursing Research*. 2012; 61:181–187. [PubMed: 22551992]
 30. Polit DF, Beck CT. The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*. 2006; 29:489–497. [PubMed: 16977646]
 31. R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria; 2012. <http://www.R-project.org/>
 32. Rosseel, Y. lavaan: Latent Variable Analysis. R package version 0.3-1. 2010. <http://CRAN.R-project.org/package=lavaan>
 33. Rosseel Y. Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*. 2012; 48:1–36.
 34. Samaniego FJ, Reneau DM. Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*. 1994; 89:947–957.
 35. Speight J, Barendse SM. FDA guidance on patient reported outcomes. *British Medical Journal*. 2010; 340:c2921. [PubMed: 20566597]
 36. Thomas JH, Nan ML. fgui: A Method for Automatically Creating Graphical User Interfaces for Command-Line R Packages. *Journal of Statistical Software*. 2009; 30:1–14. URL <http://www.jstatsoft.org/v30/i02/>. [PubMed: 21666874]
 37. Westland JC. Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*. 2010; 9:476–487.
 38. Westland, JC. *Electronic Commerce Research and Applications*. (in press). Erratum to “Lower bounds on sample size in structural equation modeling” [Electron. Commerce Res. Appl. 9 (6) (2010) 476–487] Available online 28 June 2012
 39. Wilks, SS. *Mathematical Statistics*. New York: Wiley; 1962.

Figure 1.

Mean Square Error correlation estimates ρ in stimulation study using BID (dashed line) and classical factor analysis (solid line) when the number of items is 8. The number of experts are 1, 2, 6, and 16, and they are either unbiased ($\rho_0=0.5$, left panel) or biased ($\rho_0=0.75$, right panel). The participant sample sizes are 50 (A and B), 100 (C and D), and 200 (E and F).

Figure 2.

Mean Square Error (A) and Square of bias (B) of correlation estimates ρ in stimulation study using classical factor analysis, BID modeling with a panel of six experts containing one biased prior ($C = 1$), and BID modeling with a panel of six experts containing two biased priors ($C = 2$) when participant sample size is 50 (a), 100 (b) and 200 (c). All scenarios assume a true correlation $\rho_0=0.5$ with biased expert judgment of $\rho_0=0.75$.