Revista Colombiana de Estadística - Applied Statistics
July 2021, volume 44, issue 2, pp. 313–329
http://doi.org/10.15446/rce.v44n2.87690

Bayesian Hierarchical Factor Analysis for Efficient Estimation Across Race/Ethnicity

Estimación eficiente a través de raza y etnicidad usando análisis factorial jerárquico bayesiano

Jinxiang Hu[1,a], Lauren Clark[1,b], Peng Shi[1,c], Vincent S. Staggs[2,d], Christine Daley[3,e], Byron Gajewski[1,f]

[1]Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, USA

[2]Health Services & Outcomes Research, University of Missouri Medical Center, Kansas City, USA

[3]Department of Family Medicine, University of Kansas Medical Center, Kansas City, USA

Abstract

Patient reported outcomes are gaining more attention in patient-centered health outcomes research and quality of life studies as important indicators of clinical outcomes, especially for patients with chronic diseases. Factor analysis is ideal for measuring patient reported outcomes. If there is heterogeneity in the patient population and when sample size is small, differential item functioning and convergence issues are challenges for applying factor models. Bayesian hierarchical factor analysis can assess health disparity by assessing for differential item functioning, while avoiding convergence problems. We conducted a simulation study and used an empirical example with American Indian minorities to show that fitting a Bayesian hierarchical factor model is an optimal solution regardless of heterogeneity of population and sample size.

Key words: American Indians; Bayesian hierarchical model; Differential item functioning; Factor analysis; Health disparities; Patient reported outcomes.

[a]Ph.D. E-mail: jhu2@kumc.edu
[b]Ph.D. E-mail: lclark5@kumc.edu
[c]Ph.D. E-mail: pshi@kumc.edu
[d]Ph.D. E-mail: vstaggs@cmh.edu
[e]Ph.D. E-mail: cdaley@kumc.edu
[f]Ph.D. E-mail: bgajewski@kumc.edu

Resumen

Las repuestas reportadas por el paciente están siendo fuertemente consideradas en la investigación de respuestas de salud centradas en el paciente y en estudios de calidad de vida comos indicadores importantes de respuestas clínicas, especialmente en pacientes con enfermedades crónicas. El análisis factorial es ideal para medir respuestas reportadas por el paciente. Cuando hay heterogeneidad en la población de pacientes y el tamaño muestral es pequeño, diferencias en el funcionamiento de los ítems y problemas de convergencia plantean dificultades para aplicar modelos factoriales. El análisis factorial jerárquico Bayesiano puede evaluar disparidades de salud evaluando el funcionamiento diferencial de los ítems, mientras que evita problemas de convergencia. Hemos realizado un estudio de simulación y empleado un ejemplo empírico con minorías indígenas Americanas para mostrar que el ajuste de un modelo factorial jerárquico Bayesiano es una solución óptima sin importar la heterogeneidad de la población o el tamaño muestral.

***Palabras clave*:** Análisis factorial; Disparidades en salud; Funcionamiento diferencial de ítems; Indígena americano; Modelo jerárquico Bayesiano; Respuestas reportadas por el paciente.

1. Introduction

Patient reported outcomes (PRO) are variables reflecting the status of a patient's health condition whose values comes directly from the patient (FDA-NIH Biomarker Working Group, 2016). PROs are gaining more attention in patient-centered health outcomes research and quality of life (QOL) studies as important indicators of clinical outcomes for patients including chronic diseases (diseases that last for year or more and require ongoing treatment; Centers for Disease Control and Prevention (2021)). PROs are especially important when symptoms, functioning, and well-being are main concerns. There have been discrepancies between patient and clinician perspectives on certain disease outcomes (e.g. Basch et al., 2011; Cuijpers et al., 2010; Sanderson & Kirwan, 2009). PROs allow researchers to assess treatment effects and satisfaction with health services from patients' own perceptions of changes in symptoms, in mental, physical, and social health, and in overall QOL (Cella et al., 2010; Deshpande et al., 2011). PROs are also easier and less expensive to administer than clinician reported measures. The importance of PROs has been advocated by health regulatory authorities (Doward et al., 2010; European Medicines Agency, 2016; US Food and Drug Administration, 2009; US Food & Drug Administration, 2020). The National Institutes of Health (NIH) funded the development of the Patient-Reported Outcomes Measurement Information System (PROMIS) for purposes of clinical research across a variety of chronic diseases (National Institutes of Health, 2019). PROs often are measured using survey questionnaires with ordinal response scales, the development of which must go through rigorous testing and validation to ensure their psychometric integrity (Dawson et al., 2010; Garrard et al., 2015). Standardized PROs such as the PROMIS measures allow comparisons to be made across studies. They can

also increase precision of results, reduce burden on patients, and improve decision-making in clinical trials Frost et al. (2007), Ware Jr et al. (2004), as well as more accurately reflect long term clinical changes and assist in better individualized treatment plans and clinical decisions (Calvert et al., 2018; European Medicines Agency, 2016; Tunis et al., 2003; US Food and Drug Administration, 2009).

Because PROs typically reflect hypothetical constructs like QOL that cannot be directly observed, latent variable methods are ideal for analyzing PROs, e.g., Joreskog et al. (1979) and MacCallum & Austin (2000). Factor analysis (FA), for example, accounts for the measurement errors in the hypothetical constructs that underlie observed variables like PROs (Muthén, 2002). In factor analysis, maximum likelihood estimation (MLE) has been criticized for always leading to the rejection of the model (Marsh et al., 2009) and for its susceptibility to the identification problem (MacCallum et al., 1992). When sample size is limited, e.g. in the case of rare diseases where sample size is often 50 or less in clinical trials (Bell & Smith, 2014), MLE is fallible.

Bayesian factor analysis is an alternative FA method where the model is identified by incorporating prior information about the distribution of model parameters. Compared to maximum likelihood estimation (MLE), Bayesian estimates generally have smaller error variance but larger bias (Chaloner, 1987). Bayesian FA performs better than traditional FA with small samples (Asparouhov & Muthén, 2010; McNeish, 2016); Muthén (2010); Muthén & Asparouhov (2012); Scheines et al. (1999) and also has better convergence properties than MLE in confirmatory FA (Hoofs et al., 2018). Bayesian estimation in FA was discussed by Press (2009) and Lee (2007). Bayesian analysis (Gelman et al., 2006; Kruschke, 2011; Lynch, 2007)) became popular due to the success of the new Markov Chain Monte Carlo (MCMC) computational methods (Gamerman & Lopes, 2006; Gilks et al., 1996). The MCMC makes use of the Gibbs sampling to draw random samples from posterior distribution. When the priors carry inaccurate information about the parameters, Bayesian estimates should be processed with caution (Samaniego & Reneau, 1994; Jiang et al., 2014). Differential item functioning (DIF) occurs when an item has different likelihood of being endorsed or answered correctly by individuals from different groups (e.g. age or racial groups) even though the individuals have the same factor score on the construct the item is designed to measure (Drasgow, 1987). DIF analysis is critical for PROs, as health disparities due to race/ethnicity may affect PRO results when minority groups are involved in a study, and DIF can interfere with our ability to quantify these disparities. For example, a QOL item may function differentially for a certain minority group by eliciting overly optimistic responses from members of that group relative to non-minority individuals with the same level of QOL, perhaps because a word in the item has a different shade of meaning in the minority group's culture. If members of the group tend to have lower QOL due to some disparity in access to health services, the systematic bias in the item could make it more difficult to detect the effects of this disparity.

Several psychometric methods have been proposed for DIF analysis (e.g. Mantel-Haenszel [MH], Wainer & Braun (1988); the item response theory-likelihood ratio method [IRT-LR], Thissen (1988); logistic regression [LR],

Swaminathan & Rogers (1990); and the multiple indicators-multiple causes method [MIMIC], Jöreskog & Goldberger (1975)). But these methods either do not account for measurement error (e.g. CTT methods including LR and MH), or require larger sample sizes (LRT-LR, MIMIC).

We propose a Bayesian Hierarchical (Gelman et al., 2013) Factor Analysis (BH-FA) method for assessing DIF. Bayesian hierarchical modeling can be extended to multilevel data (Gelman et al., 2012), which makes BH-FA appealing in that potential DIF may exist when there are multiple racial/ethnic groups present in the sample. Bayesian hierarchical modeling has been widely used in research studies in various fields (e.g. Kemp et al., 2007; Gajewski et al., 2019). In most cases the hierarchical model has been applied to account for clustering within sites or providers (Kwok & Lewis, 2011), in contexts where clustering is viewed as a nuisance. Here we examine a context where the clusters are racial/ethnic groups, and we want to make inferences regarding specific clusters. BH-FA could be especially valuable when the population is small and/or heterogeneous. To the best of our knowledge, no literature has examined the advantages of the Bayesian hierarchical approach for estimating differential item functioning across racial/ethnic groups in factor analysis. Our purpose is to fill this gap in the literature, by conducting a simulation study. We want to examine the advantages of BH-FA in assessing DIF in an instrument, and also accommodate small sample size and convergence problem in latent factor analysis.

2. Method

Our motivation comes from a study where the Patient Assessment of Mammography Services (PAMS) survey was administered to a sample that included ethnic minority women (Ndikum-Moffor et al., 2013). The PAMS instrument was developed by researchers in a Midwestern academic medical center to assess patients' satisfaction with mammography services. The full PAMS instrument comprises 20 items designed to load on four factors, and the PAMS short form comprises 7 items designed to measure one factor (Engelman et al., 2010). Participants came from four racial/ethnic groups: White, Hispanic, African American, and American Indian.

Table 1: PAMS questions.

Item 1	Overall, how would you rate the appointment scheduling process?
Item 2	Overall, how would you rate your level of comfort during the exam?
Item 3	Overall, how would you rate the comfort of the mammography facility?
Item 4	Overall, how would you rate the convenience of the mammography facility?
Item 5	Overall, how would you rate the mammography technologist?
Item 6	Overall, how would you rate the way you received your mammography test results?
Item 7	Overall, how would you rate your entire mammography experience?

Motivated by the PAMS study, we want to use a simulation study to examine the differences and similarities between three different possible models

for analyzing the data: a homogeneous model, an independent model, and the BH-FA model. The simulation study also allowed us to calibrate the BH-FA model for application to the PAMS dataset. We then compare the models on the original dataset. Inspired by the PAMS study, we simulated data for a one factor model with 7 items and 4 race groups (Figure 1). Data were simulated from two versions of the true model: one with group differences (independent model) and one without (homogeneous model). Our purpose was to properly calibrate the priors in the BH-FA model through simulation.

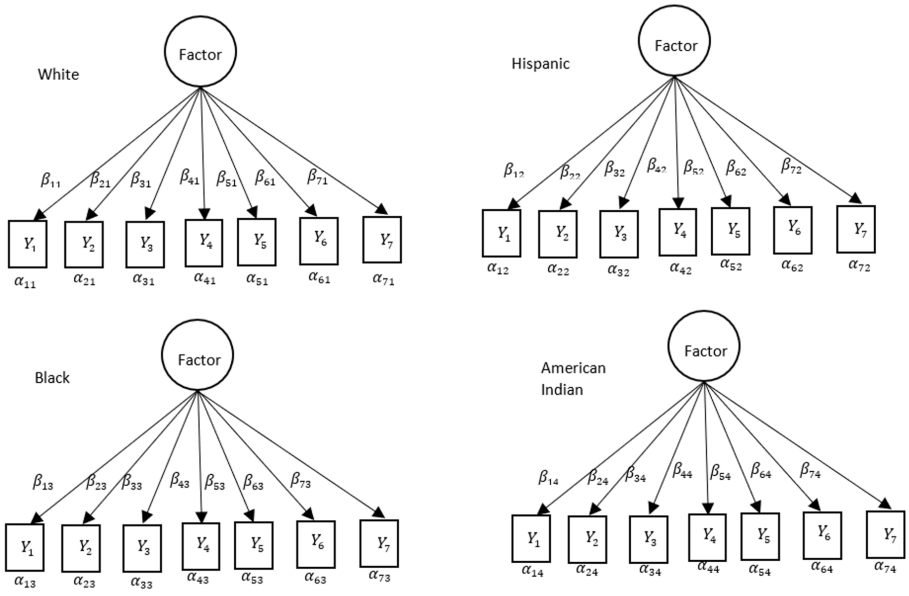


FIGURE 1: CFA model.

Data were generated within Mplus. Continuous item scores were simulated from a multivariate normal distribution. Simulation conditions included: sample size (50, 100, 500 in each racial/ethnic group), intercept difference between groups (0 indicates no group difference, 1 indicates a group difference), and slope that represents the correlation between the items and the latent factor (.2, .4, .6, .8). Additionally, in order to examine the effect of combined intercept and slope difference, we added a condition to the slope parameter when intercept difference is 1 to represent each group has both different intercepts and different slopes (.2, .4, .6, .8 in each group respectively). Altogether there are $3 \times 2 \times 4 + 3 = 27$ different conditions simulated. Each condition was iterated 200 times, and altogether there were 5,400 datasets simulated. Each dataset was analyzed using the homogeneous model, the independent model, and the hierarchical model. Altogether $5,400 \times 3 = 16,200$ analyses were performed. The homogeneous and independent models were fitted in Mplus, and the hierarchical model was fitted in WinBUGS.

2.1. Confirmatory Factor Analysis Model

We simulated data from the following model:

$$\begin{aligned} y_{ijk} &= \alpha_{jk} + \beta_{jk} f_i + \epsilon_{ijk} \\ i = 1, \ldots, N; j &= 1, \ldots, 7; k = 1, \ldots, 4 \end{aligned} \tag{1}$$

where y_{ijk} denotes the observed response for the ith person to the jth item, k denotes the group (e.g., race) to which the ith person belongs, α_{jk} is the mean response to the jth item in the kth group, f_i is the standard normal latent factor score for the ith person, f_i $N(0,1)$, β_{jk} is the factor loading representing the strength of the relationship between f_i and the observed response to the jth item for participants in the kth group, and ϵ_{ijk} is the measurement error or residual for the ijkth response, ϵ_{ijk} $N(0, \sigma^2_{jk})$. Thus the distribution of the responses is y_{ijk} $N(\alpha_{jk}, \beta^2_{jk} + \sigma^2_{jk})$. The likelihood function to estimate the model is:

$$\begin{aligned} L(\alpha_{jk}, \beta_{jk} \mid y_{1jk}, \ldots, y_{Njk}) &= \prod_{i=1}^{N} p(y_{ijk} \mid \alpha_{jk}, \beta_{jk}) \\ &= \left(2\pi(\beta^2_{jk} + \sigma^2_{jk})\right)^{-\frac{n}{2}} \\ &\quad \times \exp\left\{ \left(-\frac{1}{2(\beta^2_{jk} + \sigma^2_{jk}))} \right) \sum_{i=1}^{N} (y_{ijk} - \alpha_{jk})^2 \right\} \end{aligned} \tag{2}$$

2.2. Homogenous Model

In this model, α_j and β_j are estimated for each item without regard to group information. In other words, we assume there is no DIF: $\alpha_j = \alpha_{j1} = \cdots = \alpha_{jk}$ and $\beta_j = \beta_{j1} = \cdots = \beta_{jk}$. Bayesian factors analysis is sensitive to the choice of priors, since there is substantial uncertainty about the true values of the factor loadings and error variances a priori (Ghosh & Dunson, 2008).

The homogeneous model was fitted in Mplus. We used the default, non-informative priors in Mplus. The priors used for α_j, β_j and σ^2_j were:

$$\alpha_j \sim N(0, \infty), \beta_j \sim N(0, \infty), \sigma^2_j \sim IG(-1, 0)$$

2.3. Independent Model

The independent model is more flexible in that α_{jk} and β_{jk} are allowed to vary across groups, as well as items, and are estimated for each item and independently for each group, thereby allowing for DIF. The independent model was fitted in Mplus, again with default, non-informative priors. The priors used for α_{jk}, β_{jk} and σ^2_{jk} were:

$$\alpha_{jk} \sim N(0, \infty), \beta_{jk} \sim N(0, \infty), \sigma^2_{jk} \sim IG(-1, 0)$$

2.4. Hierarchical Model (BH-FA)

In the BH-FA model, α_{jk} and β_{jk} are estimated for each item and for each group as in the independent model, but here the parameters for the different groups are not independent, as they are related to each other through shared hyperparameters. In BH-FA we assign common hyperpriors to the parameters in the intercept and slope priors. The priors used in the hierarchical models were:

$$\alpha_{jk} \sim N(\mu_\alpha, \sigma_\alpha^2), \beta_{jk} \sim N^+(\mu_\beta, \sigma_\beta^2), \frac{1}{\sigma_{jk}^2} \sim Unif(0, 10)$$

Additionally, each of the hyperparameters was assigned its own hyperprior:

$$\mu_\alpha \sim N(3, 1), \frac{1}{\sigma_\alpha^2} \sim Unif(0, 10), \mu_\beta \sim N^+(1, 100), \frac{1}{\sigma_\beta^2} \sim Unif(0, 10)$$

The small-scale parameters in the hyperpriors were chosen to help shrink the inverse variance parameters toward zero, thus yielding vague priors for alpha and beta (Gelman et al., 2006). The hierarchical model was fitted within WinBUGS (Lunn et al., 2000) through R2WinBUGS (Sturtz et al., 2005) because WinBUGS is technically more flexible Muthén & Asparouhov (2012), and can accommodate hyperpriors for the intercept and slope priors. We used two parallel chains with a burn-in of 2000 iterations.

2.5. Evaluation of the Model

For all three models, we first evaluated bias for each parameter estimated under each condition. Bias is reported as a measure for recovery of true item parameters and can be calculated as:

$$Bias = \frac{\sum_{i=1}^{N}(\hat{X}_i - X_i)}{N} \tag{3}$$

where X_i indicates a parameter in the model. The root mean squared error (RMSE) was calculated for each condition as well for evaluation of model fit. For the homogeneous model, overall RMSE was calculated for the intercept parameter and the slope parameter. For the independent model and the hierarchical model, sub-group RMSEs based on racial/ethnic group were calculated separately for the intercept parameter and the slope parameter respectively. RMSE represents the square root of the squared difference between an estimated parameter and its true value. A smaller value for RMSE indicates the model fits the data better. RMSE can be calculated with the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{X}_i - X_i)^2}{N}} \tag{4}$$

3. Results

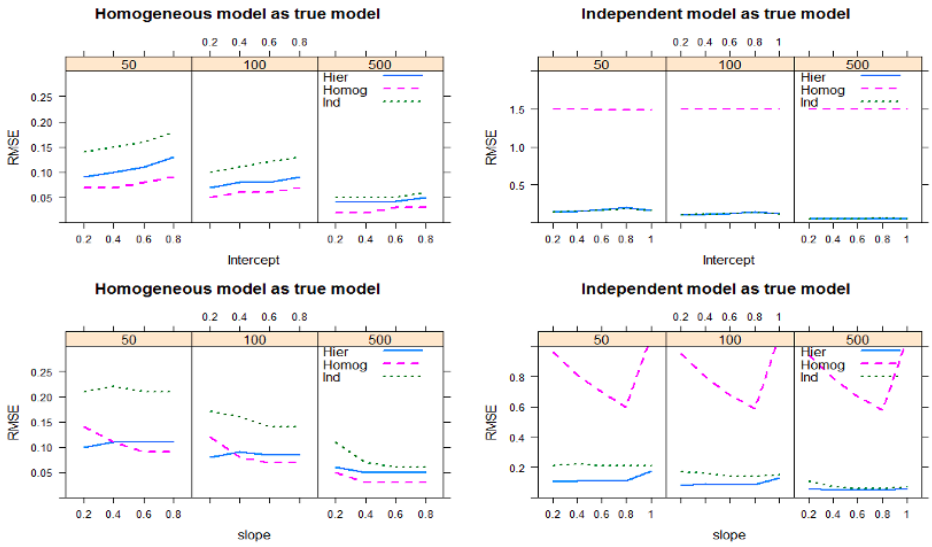
The parameter recovery was in general very good, both when the true model was the homogeneous model (intercept/α difference = 0) and when it was the independent model (intercept/α difference = 1) (Table 1). Bias of the intercept parameter estimates was only 0 or $-.01$, and bias in slope parameter estimates ranged from 0 to 0.06 in either direction. Bias seem to be larger when slope was small at .2. Overall, the bias was virtually zero when the slope was larger than .2. Sample size did not have a substantial impact on the bias for either the intercept parameter α or the slope parameter β.

Table 2: Bias (parameter recovery) in the simulation study.

Group n	α_T difference	β_T	α	β
50	0	0.2	0.00	-0.04
		0.4	0.00	0.00
		0.6	0.00	0.01
		0.8	-0.01	0.01
	1	0.2	0.00	-0.02
		0.4	0.00	-0.01
		0.6	0.00	0.03
		0.8	-0.01	0.06
		.2 .4 .6 .8	-0.01	0.02
100	0	0.2	0.00	-0.03
		0.4	0.00	0.00
		0.6	0.00	0.00
		0.8	0.00	0.01
	1	0.2	0.00	-0.04
		0.4	0.00	-0.01
		0.6	0.00	0.02
		0.8	0.00	0.03
		.2 .4 .6 .8	0.00	0.00
500	0	0.2	0.00	-0.01
		0.4	0.00	0.00
		0.6	0.00	0.00
		0.8	0.00	0.00
	1	0.2	0.00	-0.03
		0.4	0.00	0.00
		0.6	0.00	0.00
		0.8	0.00	0.00
		.2 .4 .6 .8	0.00	-0.01

Note: α_T indicates true intercept parameter;
β_T indicates true slope parameter.

In Figure 2, RMSE values under different sample sizes are plotted against the intercept/α parameter (top panel) and the slope/β parameter (bottom panel) when the true model was the homogeneous model (intercept/α difference = 0, left panel) or the independent model (intercept/α difference=1, right panel). RMSE values are shown for all three estimating models: the homogeneous model (pink dashed line), the independent model (green dashed line), and the BH-FA model (blue solid line).



Note: Homogeneous model as true model (intercept difference = 0); Independent model as true model (intercept difference = 1); Slope = M indicates slope is different for each group (.2, .4, .6, .8); homog = homogeneous model, Ind = independent model, hier = BH-FA model.

FIGURE 2: RMSE for simulated data.

When the homogeneous model was the true model (no DIF/group difference), fitting a homogenous model was best in accuracy of the intercept and the slope, followed by the BH-FA model and then the independent model. RMSE decreased as sample size increased. It is worth noting that the BH-FA model outperformed even the homogeneous model (the true model) in terms of accuracy of the slope estimation when the slope was relatively flat at .2 with sample size of 100 or less.

On the other hand, when the true model was the independent model (presence of DIF/group difference), the BH-FA model fit as well as or better than the independent model. The homogeneous model performance was very poor under this condition. When there were both intercept and slope differences (DIF) between groups, the independent model and the BH-FA model fit equally well, while the homogeneous model fit very poorly. Sample size did not have a large effect on RMSE estimates when the true model was the independent model.

3.1. Empirical Example: PAMS Study

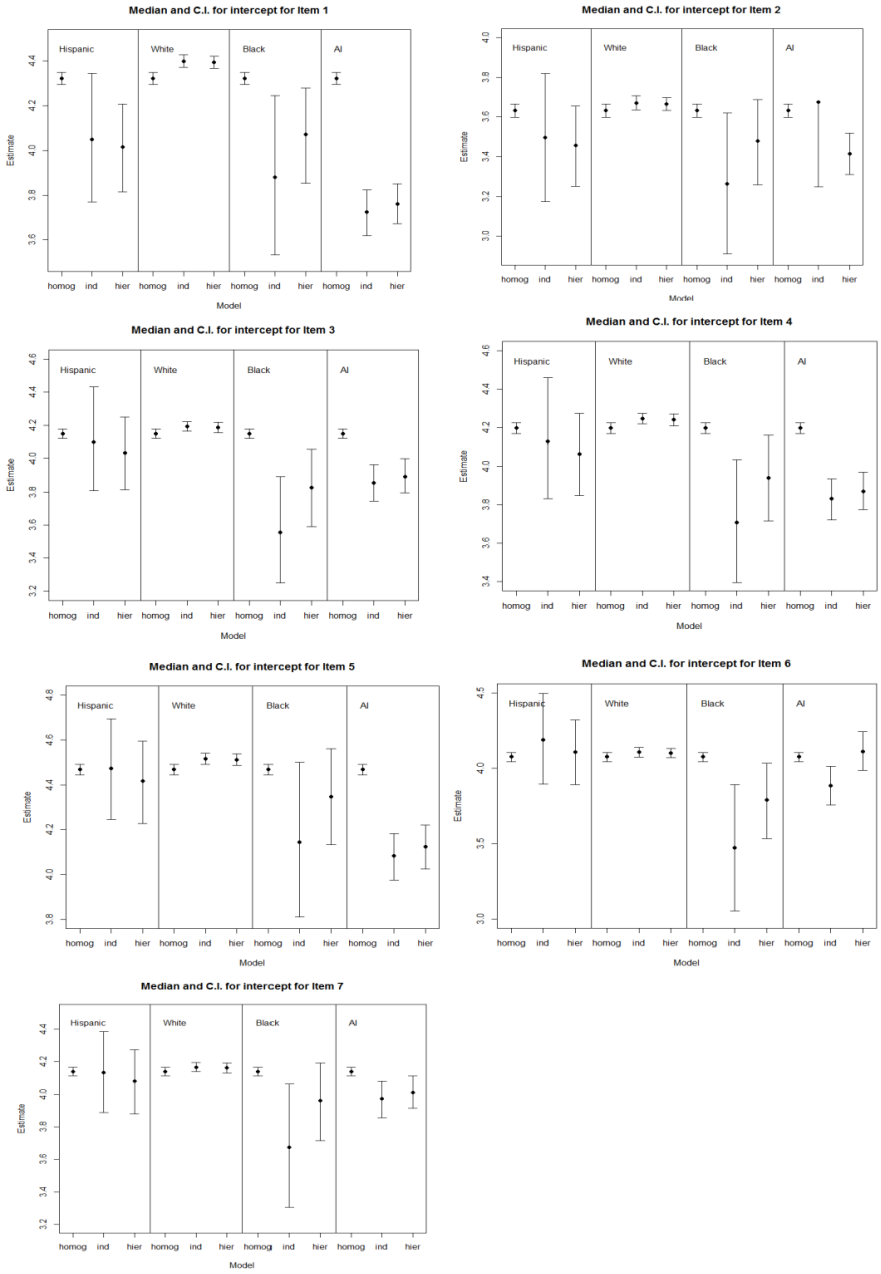
We fit the homogeneous model, the independent model, and the BH-FA model to the PAMS data with the priors in the simulation study. We plotted the median and the equal-tailed 95% credibility interval (CI) for the intercept parameter (Figure 3) and the slope parameter (Figure 4) estimates for the 7 PAMS questions, stratified by group (White, Hispanic, African American, and American Indian). In the PAMS case, it was clear that the population was heterogeneous with 4 racial/ethnic groups. When group sample size was large, all three models produced similar estimates for the intercept and the slope, as well as their CIs, as seen in the White group. On the other hand, when sample size was small, as in the Hispanic and the AI groups, the CIs were much larger under the independent and the BH-FA models.

In the PAMS case, the homogeneous model produced falsely similar estimates for all four groups, as well as overly narrow credibility intervals. This was not surprising as the homogeneous model ignored the differences across groups. When we used the independent model or the BH-FA model, the estimates for the intercept and the slope were different for each group, and the BH-FA estimates were very similar to the independent model estimates. The pattern was the same with the estimates of the CI of the parameters, the BH-FA model produced narrower CI. Specifically, the CI was the smallest for the White group due to its large sample size, and largest for the Hispanics and the AI because of small sample size. Between the independent model and the BH-FA model, the BH-FA model had narrower CI due to the hyper parameter imposed. This was consistent with our simulation study findings, and suggested an advantage of using the BH-FA model when data are heterogeneous across groups and DIF may be present.

4. Conclusions and Future Work

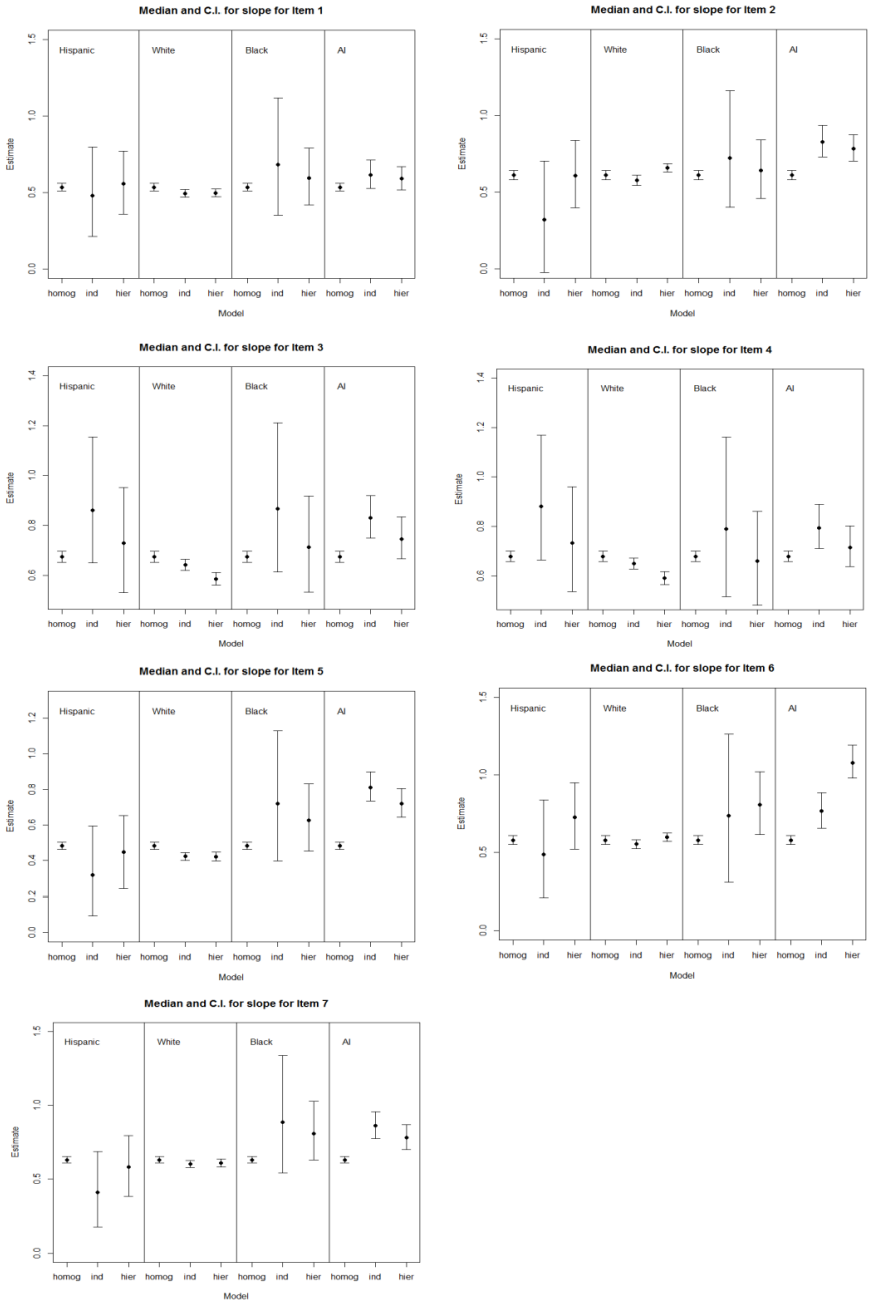
PROs are important measures in health and clinical research. And it is very common to have data collected from multiple sites or nested multilevel data. Under some situations, such as in studies of a rare disease, small sample size is unavoidable. Under these circumstances, fitting the BH-FA model is an optimal solution because the BH-FA model fits the data well regardless of the data structure. Motivated by the PAMS study we used a simulation study to calibrate the hyper parameters in the BH-FA model, and assessed its performance with and without the presence of DIF in the PRO instrument used. Our conclusion is the BH-FA can facilitate the potential DIF effect, and thus provide better PRO score estimates. This, in turn, may improve our ability to identify health disparities with a note of statistical methods can only be helpful when the content validity of the instrument has been established.

The study has its limitations. We only examined the RMSE for model evaluation and did not examine effects of missing data. Given our findings, future work can explore the BH-FA for use in PRO instrument development and sample size smaller than 50 such as in the case of rare diseases. If necessary, informative priors can be added to the model with well knowledge about the population parameter.



Note: homog = homogeneous model, Ind = independent model, hier = BH-FA model, CI = credibility interval, AI = American Indian

FIGURE 3: RMSE median and credibility interval for the PAMS items intercepts estimated with the homogeneous model, independent model and the hierarchical model.



Note: homog = homogeneous model, Ind = independent model, hier = BH-FA model, CI = credibility interval, AI = American Indian

FIGURE 4: RMSE median and credibility interval for the PAMS items slopes estimated with the homogeneous model, independent model and the hierarchical model.

Acknowledgements

This work was supported by the University of Kansas Cancer Center (P30 CA168524), by a National Institute on Minority Health and Health Disparities grant (P20MD004805) awarded to Center for American Indian Community Health, and by a Clinical and Translational Science Institute TSA grant from the National Center for Advancing Translational Sciences awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (UL1TR002366). We also would like to thank the participants who participated in the Patient Assessment of Mammography Services survey. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS.

[Received: December 2020 — Accepted: March 2021]

References

Asparouhov, T. & Muthén, B. (2010), 'Bayesian analysis of latent variable models using mplus'.

Basch, E., Bennett, A. & Pietanza, M. C. (2011), 'Use of patient-reported outcomes to improve the predictive accuracy of clinician-reported adverse events'.

Bell, S. A. & Smith, C. T. (2014), 'A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of clinicaltrials. gov', *Orphanet Journal of Rare Diseases* **9**(1), 1–11.

Calvert, M., Kyte, D., Mercieca-Bebber, R., Slade, A., Chan, A.-W., King, M. T., Hunn, A., Bottomley, A., Regnault, A., Ells, C. et al. (2018), 'Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the spirit-pro extension', *Jama* **319**(5), 483–494.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S. et al. (2010), 'The patient-reported outcomes measurement information system (promis) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008', *Journal of Clinical Epidemiology* **63**(11), 1179–1194.

Centers for Disease Control and Prevention (2021), 'About chronic diseases'. https://www.cdc.gov/chronicdisease/about/index.htm

Chaloner, K. (1987), 'A bayesian approach to the estimation of variance components for the unbalanced one-way random model', *Technometrics* **29**(3), 323–337.

Cuijpers, P., Li, J., Hofmann, S. G. & Andersson, G. (2010), 'Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy

research on depression: a meta-analysis', *Clinical Psychology Review* **30**(6), 768–778.

Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C. & Carr, A. J. (2010), 'The routine use of patient reported outcome measures in healthcare settings', *BMJ* **340**.

Deshpande, P. R., Rajan, S., Sudeepthi, B. L. & Nazir, C. A. (2011), 'Patient-reported outcomes: a new era in clinical research', *Perspectives in clinical research* **2**(4), 137.

Doward, L. C., Gnanasakthy, A. & Baker, M. G. (2010), 'Patient reported outcomes: looking beyond the label claim', *Health and quality of life outcomes* **8**(1), 1–9.

Drasgow, F. (1987), 'Study of the measurement bias of two standardized psychological tests.', *Journal of Applied psychology* **72**(1), 19.

Engelman, K. K., Daley, C. M., Gajewski, B. J., Ndikum-Moffor, F., Faseru, B., Braiuca, S., Joseph, S., Ellerbeck, E. F. & Greiner, K. A. (2010), 'An assessment of american indian women's mammography experiences', *BMC women's health* **10**(1), 1–13.

European Medicines Agency (2016), 'Guideline on the evaluation of anticancer medicinal products in man: The use of patient-reported outcome (pro) measures in oncology studies'.
https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf

FDA-NIH Biomarker Working Group (2016), 'Best (biomarkers, endpoints, and other tools) resource [internet]'.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D. & Group, M. P.-R. O. C. M. (2007), 'What is sufficient evidence for the reliability and validity of patient-reported outcome measures?', *Value in Health* **10**, S94–S105.

Gajewski, B. J., Meinzer, C., Berry, S. M., Rockswold, G. L., Barsan, W. G., Korley, F. K. & Martin, R. (2019), 'Bayesian hierarchical emax model for dose-response in early phase efficacy clinical trials', *Statistics in Medicine* **38**(17), 3123–3138.

Gamerman, D. & Lopes, H. F. (2006), *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, CRC Press.

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), 1–14.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Gelman, A., Hill, J. & Yajima, M. (2012), 'Why we (usually) don't have to worry about multiple comparisons', *Journal of Research on Educational Effectiveness* **5**(2), 189–211.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian Analysis* **1**(3), 515–534.

Ghosh, J. & Dunson, D. B. (2008), Bayesian model selection in factor analytic models, *in* 'Random effect and latent variable model selection', Springer, pp. 151–163.

Gilks, W., Richardson, S. & Spiegelhalter, D. (1996), 'Markov chain monte carlo in practice chapman and hallcrc'.

Hoofs, H., van de Schoot, R., Jansen, N. W. & Kant, I. (2018), 'Evaluating model fit in bayesian confirmatory factor analysis with large samples: Simulation study introducing the brmsea', *Educational and Psychological Measurement* **78**(4), 537–568.

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**(4), 296–310.

Joreskog, K. G., Dag, S. & Magidson, J. (1979), *Advances in factor analysis and structural equation models*, Abt books.

Jöreskog, K. G. & Goldberger, A. S. (1975), 'Estimation of a model with multiple indicators and multiple causes of a single latent variable', *journal of the American Statistical Association* **70**(351a), 631–639.

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2007), 'Learning overhypotheses with hierarchical bayesian models', *Developmental science* **10**(3), 307–321.

Kruschke, J. (2011), Tutorial: Doing bayesian data analysis with r and bugs, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 33.

Kwok, H. & Lewis, R. J. (2011), 'Bayesian hierarchical modeling and the integration of heterogeneous information on the effectiveness of cardiovascular therapies', *Circulation: Cardiovascular Quality and Outcomes* **4**(6), 657–666.

Lee, S.-Y. (2007), *Structural equation modeling: A Bayesian approach*, Vol. 711, John Wiley & Sons.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000), 'Winbugs-a bayesian modelling framework: concepts, structure, and extensibility', *Statistics and computing* **10**(4), 325–337.

Lynch, S. M. (2007), *Introduction to applied Bayesian statistics and estimation for social scientists*, Springer Science & Business Media.

MacCallum, R. C. & Austin, J. T. (2000), 'Applications of structural equation modeling in psychological research', *Annual review of psychology* **51**(1), 201–226.

MacCallum, R. C., Roznowski, M. & Necowitz, L. B. (1992), 'Model modifications in covariance structure analysis: the problem of capitalization on chance.', *Psychological bulletin* **111**(3), 490.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. & Trautwein, U. (2009), 'Exploratory structural equation modeling, integrating cfa and efa: Application to students' evaluations of university teaching', *Structural equation modeling: A multidisciplinary journal* **16**(3), 439–476.

McNeish, D. (2016), 'On using bayesian methods to address small sample problems', *Structural Equation Modeling: A Multidisciplinary Journal* **23**(5), 750–773.

Muthén, B. (2010), 'Bayesian analysis in mplus: A brief introduction'.

Muthén, B. & Asparouhov, T. (2012), 'Bayesian structural equation modeling: a more flexible representation of substantive theory.', *Psychological methods* **17**(3), 313.

Muthén, B. O. (2002), 'Beyond sem: General latent variable modeling', *Behaviormetrika* **29**(1), 81–117.

National Institutes of Health (2019), 'Patient-reported outcomes measurement information system (promis)'. https://commonfund.nih.gov/promis/index

Ndikum-Moffor, F. M., Braiuca, S., Daley, C. M., Gajewski, B. J. & Engelman, K. K. (2013), 'Assessment of mammography experiences and satisfaction among american indian/alaska native women', *Women's Health Issues* **23**(6), e395–e402.

Press, S. J. (2009), *Subjective and objective Bayesian statistics: Principles, models, and applications*, Vol. 590, John Wiley & Sons.

Samaniego, F. J. & Reneau, D. M. (1994), 'Toward a reconciliation of the bayesian and frequentist approaches to point estimation', *Journal of the American Statistical Association* **89**(427), 947–957.

Sanderson, T. & Kirwan, J. (2009), 'Patient-reported outcomes for arthritis: Time to focus on personal life impact measures?'.

Scheines, R., Hoijtink, H. & Boomsma, A. (1999), 'Bayesian estimation and testing of structural equation models', *Psychometrika* **64**(1), 37–52.

Sturtz, S., Ligges, U. & Gelman, A. E. (2005), 'R2winbugs: a package for running winbugs from r'.

Swaminathan, H. & Rogers, H. J. (1990), 'Detecting differential item functioning using logistic regression procedures', *Journal of Educational measurement* **27**(4), 361–370.

Thissen, D. (1988), 'Use of item response theory in the study of group differences in trace lines', *Test validity* .

Tunis, S. R., Stryer, D. B. & Clancy, C. M. (2003), 'Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy', *Jama* **290**(12), 1624–1632.

US Food & Drug Administration (2020), 'Fda patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making'.
https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical

US Food and Drug Administration (2009), 'Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims'.
https://www.fda.gov/media/77832/download

Wainer, H. & Braun, H. (1988), 'Differential item performance and the mantel-haenszel procedure', *Test Validity* pp. 129–145.

Ware Jr, J., Kosinski, M. & Bjorner, J. (2004), 'Item banking and the improvement of health status measures', *Quality of Life* **2**, 2–5.