A Multidimensional Zero-Inflated Graded Response Model for Ordinal Symptom Data: Online Supplemental Materials

Bayesian Inference

We applied the proposed method in the Bayesian framework for two main reasons: (1) Bayesian inference, and (2) MCMC estimation advantages. In terms of inference, the Bayesian framework presents multiple advantages, such as direct inference for the parameters (θ) rather than indirect inference based on a null hypothesis. The inclusion of priors supports the process of cumulative science by allowing us to take into account the information gained from previous research and help guard against inadmissible parameter estimates. And compared to the frequentist confidence interval, the Bayesian credible interval has a more intuitive interpretation that does not depend on the long-run behavior of the estimator (Kruschke, 2010). In terms of estimation, MCMC has been shown to be robust in estimating complex models that can be difficult to estimate with frequentist techniques (such as maximum likelihood). For example, computational challenges due to multidimensionality and cell sparseness related to small sample size can more easily be overcome within the Bayesian framework (Huo et al., 2016). And finally, the proposed model is relatively straightforward to implement in the Bayesian program Stan (Carpenter et al., 2017). For a more thorough treatment of the comparison between Bayesian and frequentist inference, the reader is referred to Efron (1986), Matthews (2001), Wagenmakers, Lee, Lodewyckx, and Iverson (2008), and Kruschke (2010).

Prior Sensitivity Analysis

For the empirical analysis of the PHQ-9, we performed a prior sensitivity analysis to evaluate the influence that the selected priors have on the model, allowing us to be confident that the priors are not leading the posterior distribution. Three sets of priors were tested: (1) the priors described in the previous section, which are weakly informative priors that constrain the parameter space to plausible values for a

particular model, (2) stricter priors with the same distributions as previously described, but with the variability cut in half, and (3) more diffuse priors with the same distributions as previously described, but with the variability doubled. We compared these three sets of priors according to three criteria: the out of sample expected log predictive accuracy (LOOIC), the point estimates (mean of the posterior distributions) for the MZI-GRM parameters, and the width (standard deviation) of the posterior distributions for the MZI-GRM parameters. These criteria allow us to evaluate prior sensitivity in terms of out of sample predictive accuracy, point estimate influence, and the effect on inference precision.

In comparing the out-of-sample predictive accuracy (Table S1), we found no consistent differences among the three sets of priors. The ratios of $\Delta ELPD$ to the SE of the difference are small, with the largest one being -1.2. The model with weakly informative priors shows a slightly better LOOIC. In terms of predictive accuracy, all sets of priors show functionally equivalent results. Next, we looked at differences in point estimates and variability among the three sets of priors. We compare the weakly informative priors to the other two sets of priors, as the weakly informative priors are the ones used in the present research. Table S2 presents the posterior distribution comparisons. For the point estimates, all mean differences are negligible, ranging between 0.000 and 0.040. To compare the variability between posterior distributions, we calculated the ratios between the posterior standard deviations for each group of parameters. The results indicate that the width of the posteriors are functionally the same across priors, with ratios ranging between 0.994 and 1.190. These results are not surprising, as the weakly informative priors are determined based on likely parameter values rather than sample-dependent characteristics. Further, because the sample size is large ($N = 3000$), the weight of the priors in the posterior is attenuated by the sample.

Posterior Predictive Checks

Figure S1 shows graphical displays of posterior predictive checks for the MZI-GRM[1]. The first three rows depict the comparison of the replicated and observed response frequencies for each item. As can be seen from these plots, the replicated frequencies are highly consistent with the observed frequencies for all items, suggesting that the model predicts the observed frequencies well. The model is also able to recover the observed summed score distribution fairly well, as shown in the bottom row of Figure S1. With the exception of a summed score of 1, which is slightly overpredicted by the model, the observed and posterior predictive median response frequencies are very similar. All observed responses fall between the dotted lines, which represent the 5th and 95th quantiles of the posterior predictive response frequencies.

Trace Lines for the MZI-GRM

The item trace lines for the MZI-GRM can be found in Figure S2. The trace line plots in the second and fourth rows show the probability of endorsing the item at all as a function of the susceptibility latent variable. The trace line plots in the first and third rows show the probability of endorsing each response category as a function of a severity latent variable. As previously described, the items about feeling down, depressed, or hopeless and thoughts of suicide or self-harm are considerably more discriminating along susceptibility than severity, whereas the items about sleep and low energy levels are much more discriminating along severity. In examining the severity trace lines for all items, it can also be observed that the third response category – "more than half the days" – is largely redundant with "several days" and is never the most probable category of endorsement. This has also been shown in previous research (Magnus & Liu, 2018).

---

[1] Graphical displays of posterior predictive checks for the alternative models can be found at https://osf.io/hvkfn/

Table S1
Out of sample predictive accuracy model comparison for prior sensitivity analysis.

Model	pD	LOOIC	ELPD	Δ ELPD (SE)	ΔELPD/ΔSE
Strict variance	3001.3	28732.4	-14366.2	–	–
Weakly informative	3001.0	28729.0	-14364.5	-1.7 (4.2)	-0.4
Diffuse variance	3003.8	28739.3	-14369.6	-5.1 (4.1)	-1.2

pD: Effective number of parameters; LOOIC: Leave-one-out information criterion;
ELPD: Expected log pointwise predictive density;
Δ ELPD: Change in ELPD from previous model to current model

Table S2
Parameter point estimate and variability comparison for the prior sensitivity analysis.

	Point Estimate		Variability	
Parameter	$W-S$	$W-D$	W/S	W/D
a_0	-0.004	-0.005	1.005	1.009
b_0	0.012	0.040	1.040	1.089
a_1	0.012	0.025	0.994	0.998
$b_{1.}$	-0.001	0.000	0.990	0.977
$r_{\theta_0\theta_1}$	0.009	0.018	1.089	1.190

$W-S$: mean difference between weakly informative and strict priors;
$W-D$: mean difference between weakly informative and diffuse priors;
W/S: Standard Deviation ratio between weakly informative and strict priors;
W/D: Standard Deviation ratio between weakly informative and diffuse priors

Figure S1. Graphical posterior predictive checks for the MZI-GRM. Top three rows: Violin plots of posterior predictive distributions of item-level response frequencies. Observed response frequencies are shown with solid black circles. Posterior predictive median response frequencies are shown with hollow triangles. Bottom row: Violin plots of posterior predictive distributions of summed score response frequencies. Posterior predictive medians are connected with dashed lines, and the 5% and 95% quantiles are shown with dotted lines.

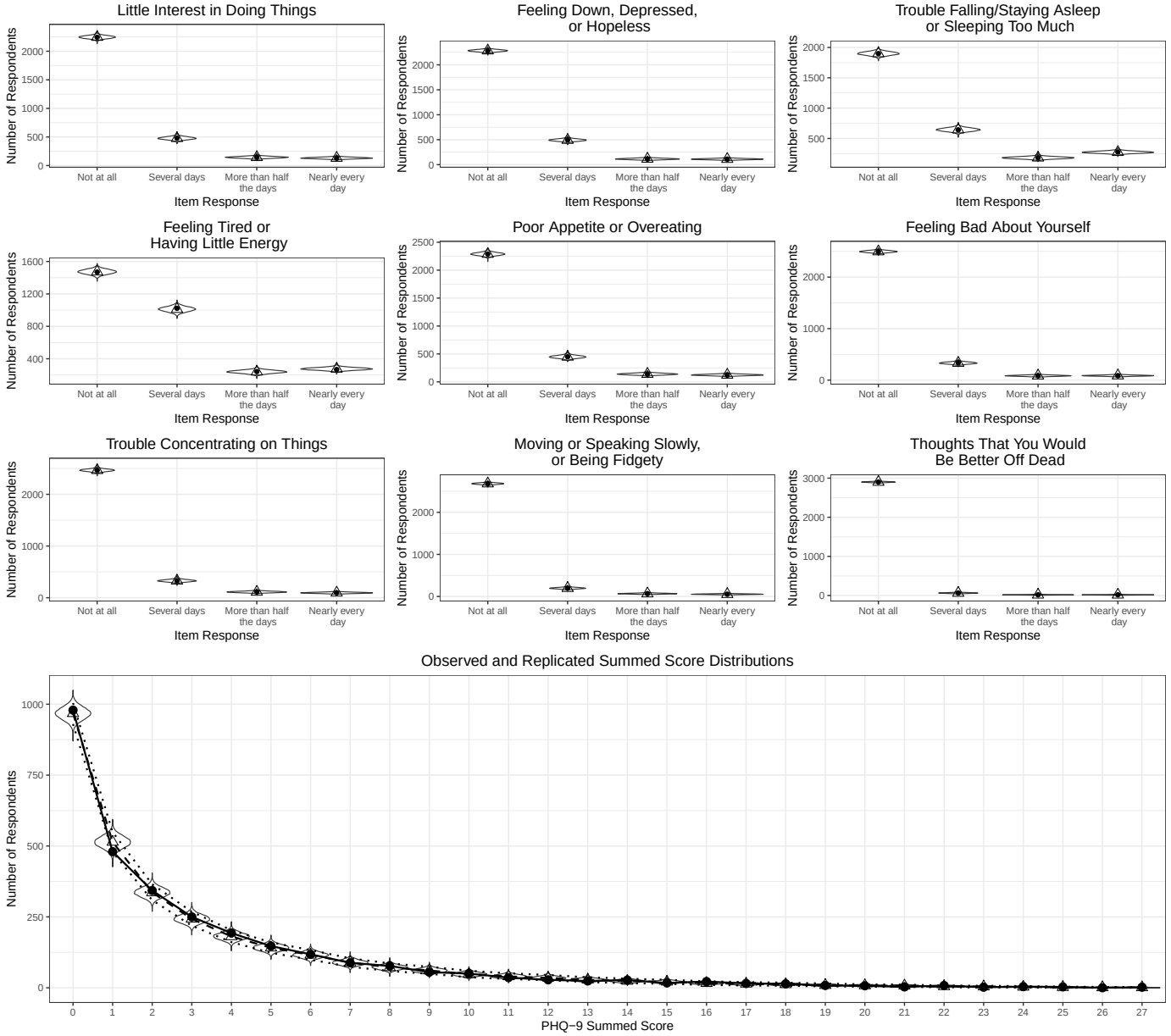
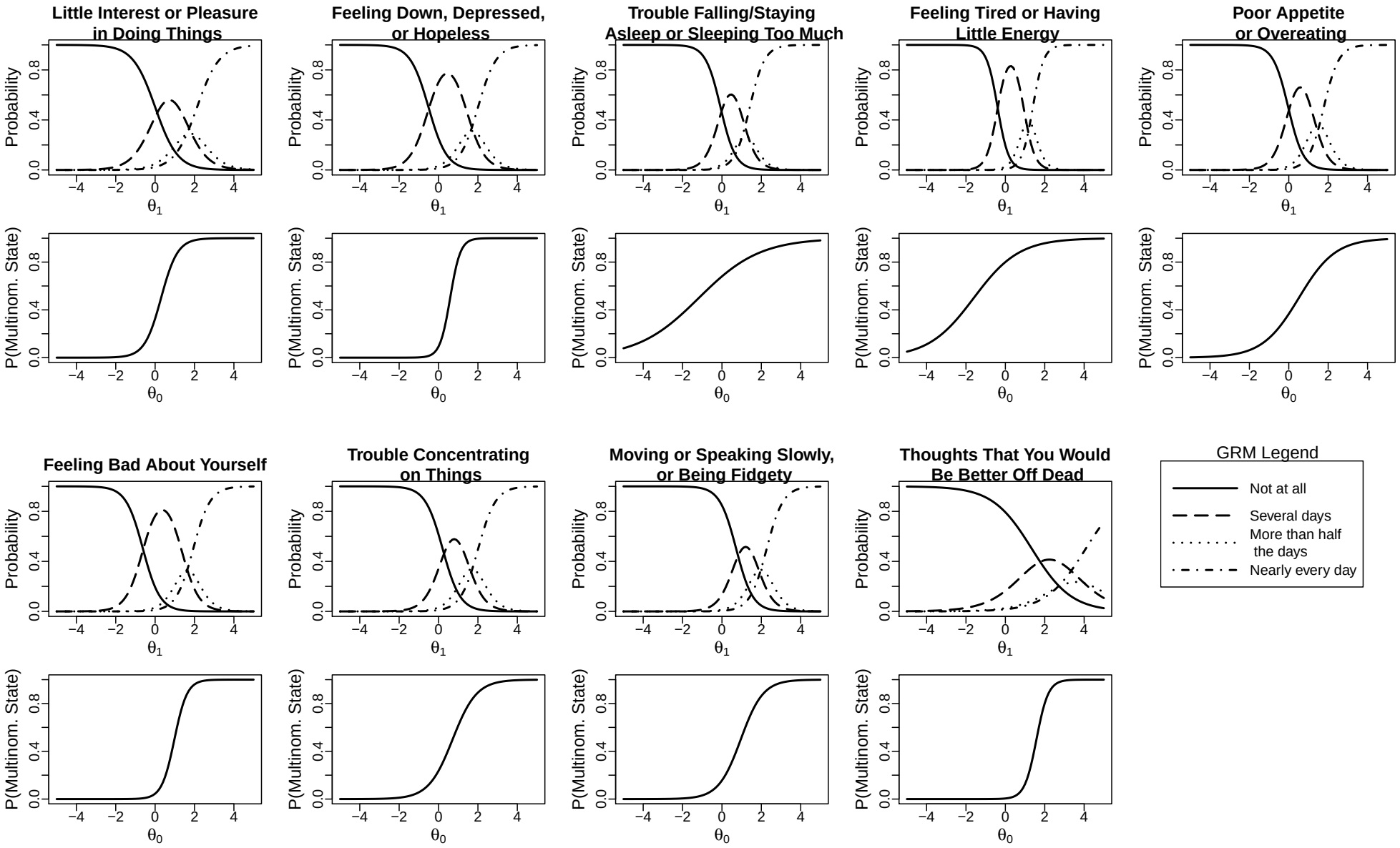


Figure S2. Trace lines for the MZI-GRM. First and third rows: Probability of endorsing a specific response category as a function of severity (θ_1); Second and fourth rows: Probability of endorsing a symptom (presence vs. absence) as a function of susceptibility (θ_0).



References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*(1), 1–5. doi: 10.2307/2683105

Huo, Y., de la Torre, J., Mun, E., Kim, S., Ray, A. E., Jiao, J., & White, H. R. (2016). A hierarchical multi-unidimensional IRT approach for analyzing sparse, multi-group data for integrative data analysis. *Psychometrika*, *80*(3), 834-855.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. doi: 10.1016/j.tics.2010.05.001

Magnus, B. E., & Liu, Y. (2018). A zero-inflated box-cox normal unipolar item response model for measuring constructs of psychopathology. *Applied Psychological Measurement*, *42*(7), 571-589. doi: 10.1177/0146621618758291

Matthews, R. A. (2001). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, *94*(1), 43–58. doi: 10.1016/S0378-3758(00)00232-9

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer. doi: 10.1007/978-0-387-09612-4$_9$