Model evaluation in the presence of categorical data:

Bayesian model checking as an alternative to traditional methods

Wes Bonifay

University of Missouri

email: bonifayw@missouri.edu

Sarah Depaoli

University of California, Merced

Published September 14, 2021.

Bonifay, W., & Depaoli, S. (2021) Model evaluation in the presence of categorical data: Bayesian model checking as an alternative to traditional methods. *Prevention Science.* https://doi.org/10.1007/s11121-021-01293-w

Peer-review process details: https://www.springer.com/journal/11121/submission-guidelines#Instructions%20for%20Authors_Editorial%20procedure

Abstract

Statistical analysis of categorical data often relies on multiway contingency tables; yet, as the number of categories and/or variables increases, the number of table cells with few (or zero) observations also increases. Unfortunately, sparse contingency tables invalidate the use of standard goodness-of-fit statistics. Limited-information fit statistics and bootstrapping procedures offer valuable solutions to this problem, but they present an additional concern in their strict reliance on the (potentially misleading) observed data. To address both of these issues, we demonstrate the Bayesian model checking technique, which yields insightful, useful, and comprehensive evaluations of specific properties of a given model. We illustrate this technique using item response data from a patient-reported psychopathology screening questionnaire, and we provide annotated R code to promote dissemination of this informative method in other prevention science modeling scenarios.

Keywords: model evaluation, Bayesian, model checking, item response theory

Model evaluation in the presence of categorical data: Bayesian model checking as an alternative to traditional methods

An important component of any statistical modeling endeavor is the thorough evaluation of the model, and in typical practice, evaluation consists primarily of testing the goodness-of-fit (GOF) of the hypothesized model to the observed data. The limitations of this approach are well documented: GOF methods are generally misunderstood and often misapplied (Roberts & Pashler, 2000) and fit indices are adversely affected by issues such as sample size (Marsh, Balla, & McDonald, 1988), non-normality (Ory & Mokhtarian, 2010), model complexity in terms of the number of parameters (Marsh & Balla, 1994) or the particular arrangement of the parameters in the model (Bonifay & Cai, 2017), and other issues (Hayduk et al., 2007).

GOF testing is even more problematic when the data are categorical. Discrete item data can be arranged in a multiway contingency table of the observed probabilities and appraised via a χ^2 test, but it is known that the asymptotic *p*-values of the χ^2 statistic are only correct when the expected frequencies in each cell of the contingency table are large (> 5 is a general guideline) (Maydeu-Olivares, 2013). Of course, the probabilities within the cells of the table must sum to one, so as the number of possible categorical data patterns increases, the expected frequencies become quite small and standard *p*-values cannot be used (Bartholomew & Tzamourani, 1999). To illustrate the scope of this sparse contingency table problem in the context of psychometric data, Maydeu-Olivares noted that when an item included four or more response categories, the classical χ^2 *p*-values became inaccurate when tests included more than five items. The same limitation applies to the G^2 likelihood ratio test.

One remedy to this issue is found in the limited-information fit statistics that have been introduced to categorical data modeling. Limited-information GOF statistics are based on the lower-order margins of the contingency table (opposed to the "full information" in each cell), usually the univariate and bivariate proportions of correct response/endorsement. Further details on GOF of categorical

data models are found throughout this Special Issue, particularly in the contribution from Cai, Chung, and Lee (in press, this volume). For a more detailed treatment of limited-information fit assessment, see Maydeu-Olivares (2013).

Another method of countering the sparse contingency table is the bootstrapping procedure (Efron, 1979). Bootstrapping is an appealing alternative to GOF testing, as it allows one to assign measures of accuracy to the estimates obtained from the observed data. Traditional (non-parametric) bootstrapping involves resampling directly from the observed data, while parametric bootstrapping involves sampling from a distribution defined by parameters estimated from the observed data. With regard to categorical data, Langeheine, Pannekoek, and van de Pol (1996) outlined how the parametric bootstrap can be applied to circumvent the contingency table sparseness and thereby obtain a distribution of some GOF statistic (rather than a single sample-dependent point estimate of GOF).

Despite skirting the sparse contingency table problem, limited-information fit and bootstrapping are both beset by a more general concern: strict reliance on the observed data. GOF statistics directly evaluate the model with regard to the available data and bootstrapping, though more informative than GOF testing, presumes that the observed data are worthy of resampling. However, if the sample data are misrepresentative of the target population (whether due to poor sampling strategies, low sample size, missing data, various breaches of data integrity, or other idiosyncrasies), then GOF may be a deceptive method of model evaluation.

An alternative to these traditional approaches can be found in the Bayesian statistical framework. Kadane (2015 in *Prevention Science*) provided a brief, yet informative overview of Bayesian inference and its potential advantages over traditional methods in the context of prevention research. As a didactic example, Kadane carried out a Bayesian analysis of data from a violence prevention study, wherein he demonstrated the benefits of allowing for distributions around the unknown parameters in a model, which is not a typical consideration of more traditional statistics. Because this work was

intended (and succeeds) as a general introduction to the Bayesian approach, it is consequently limited in scope. A key omission from Kadane's work was model evaluation. Practitioners who make use of Bayesian inference must be able to determine whether their model is "good," that is, the degree to which it accurately encapsulates the observed data and generalizes to future or otherwise unseen data. Further, and more pertinent to the aims of the present article, the Bayesian perspective offers unique insights that can supplement or dispute the findings from a traditional model evaluation.

Bayesian inference. It is generally accepted that Bayesian inference is a distinct framework that is "separate and apart" from classical inference. From the traditional perspective, Bayesian reasoning is justifiably counterintuitive, due in no small part to its underlying philosophy. The Bayesian estimation framework is also decidedly complex in that it contains statistical and estimation techniques that require a firm understanding prior to implementation. This paper is not meant to act as a primer on general Bayesian methodology, but we recommend several readings for those interested in a broader treatment of Bayesian statistics. For a gentler introduction, see Hoff (2009) or van de Schoot et al. (2014), and for a more thorough treatment see Gelman et al. (2013) or Kaplan (2014). In the current paper, we focus instead on the issue of evaluating models through the Bayesian framework, which can be beneficial to better understanding model performance. Before introducing that topic, we provide a brief overview of the statistical foundation of Bayesian estimation—namely, Bayes' theorem, which then leads into a discussion of Bayesian model evaluation.

In statistical terms, the Bayesian approach focuses on $P(\mathrm{M}|\mathrm{D})$, the probability of some hypothesized model M, given the data D. That is, Bayesian analysis treats the model as random and the data as fixed; in other words, the model may or may not be true, and any inferences must be based on the available data. Bayes' theorem (Bayes, 1764) is used to calculate $P(\mathrm{M}|\mathrm{D})$:

$$P(\mathrm{M}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{M})P(\mathrm{M})}{P(\mathrm{D})}, \tag{1}$$

where $P(\mathrm{M}|\mathrm{D})$ is known as the *posterior* probability of the model, given the data; $P(\mathrm{D}|\mathrm{M})$ is the

likelihood of the model, given the data; *P*(M) is the *prior* probability of the model; and *P*(D) is the prior probability of the data. Essentially, Eq. (1) holds that data-based inferences must consider the data and any prior beliefs about the model and/or data. Figure 1 illustrates the relationship between the likelihood, the prior, and the posterior. Specifically, the posterior distribution can be viewed as a compromise between the likelihood and the prior. Prior distributions have been a source of great controversy, largely because they have the potential to influence final model results in a substantial way. However, under cases of default software settings for priors (typically referred to as *diffuse* prior settings) and larger sample sizes, the influence of the prior is minimized.

Though much more could be said on the fundamental philosophy and implementation of Bayesian inference, our attention is on the insights that can be gained by applying Bayesian methods to evaluate categorical data models. Therefore, in the current illustration, we have opted to confine the analyses to using default prior and estimation settings in the software to provide a straightforward introduction of the topic.[1] Results should be interpreted in the context of default settings, and the reader should be aware that altering these settings in a subsequent analysis may also alter the findings. Default prior settings vary across software programs and do not uniformly impact estimates across different modeling situations. For example, if sample sizes are relatively smaller, then prior settings (whether they are default diffuse, or subjective) can have a larger impact on posterior estimates as compared to larger sample sizes (see e.g., van Erp, Mulder, & Oberski, 2018). It is also true that, as models increase in complexity (e.g., adding mixture components or multilevel structures), prior settings can impact posterior estimates in varying (and sometimes drastic) ways (see e.g., Depaoli, Yang, & Felt, 2017). Therefore, it is important for researchers to interpret the impact of prior settings in the context of the data being analyzed and the particular model being estimated. We expand on some

[1] Although we largely use default settings here, we also note that subjective prior settings are a helpful way of incorporating specific opinions, theory, or knowledge into the estimation process. We describe this element of subjectivity in more detail in the Discussion section.

methods for assessing the impact of prior settings in the Discussion section.

The Bayesian approach offers an appealing alternative to traditional model evaluation. In certain scenarios, such as linear modeling, traditional goodness-of-fit tests may be useful and informative. Further, many fit statistics, such as the χ^2 test, are easy to implement because their distribution is known (or can be approximated). Gelman, Meng, and Stern (1996) argue, however, that reliance on the traditional approach is problematic for models that impose severe restrictions on the parameters (e.g., positivity constraints) or the probability distributions (e.g., due to a strong prior) and models that do not align to the general linear model. Further, traditional methods are generally aimed at obtaining a point estimate of some parameter, while Bayes' theorem allows researchers to characterize entire distributions of uncertainty around a point estimate.

An important distinction can be made with regard to model evaluation as well. As discussed above, traditional model evaluations are typically made with reference to the observed data alone or through bootstrap resampling of the observed data. This inferential shortcoming led Gelman et al. (1996) to conclude, "It seems to us that in the context of assessing goodness-of-fit of a model for a given data set, hypothetical replications are inevitable" (p. 802). The Bayesian approach facilitates consideration of these "hypothetical replications" for *model checking*.

Model checking. Instead of gauging the degree to which a given model represents the observed data, model checking enables a researcher to investigate any feature of a model by evaluating the observed data against a reference distribution of replicated data. Gelman et al. (1996) list three ways to check a Bayesian model: by testing its sensitivity to changes in the prior probability distribution and the likelihood; by checking that inferences based on the posterior are reasonable, given the substantive research focus; and by checking that the model fits the data. We focus on this last method, though it should be noted that "fitting the data" does not refer only to global GOF, as we will explain shortly.

The goal of model checking is to evaluate a model with reference to replicated data, thereby

allowing for considerations of important issues like generalizability to future or unseen data. In that regard, model checking is conceptually similar to the traditional bootstrap technique, as noted by Gelman et al. (1996), who posited that "the posterior predictive replication appears to be the replication that the classical approach intends to address" (p. 738). This link is even more direct when considering that Rubin (1981) used the term "Bayesian bootstrap" to describe an earlier version of model checking. Yet, however similar their goals may be, model checking differs from bootstrapping in several important ways, as discussed below.

Posterior predictive model checking (PPMC), as introduced by Guttman (1967) and formally defined by Rubin (1984), involves drawing a simulated sample of the unknown parameter θ from the posterior $p(\theta|M,D)$. Given the simulated sample, a replicated dataset ($\mathrm{D}^{\mathrm{rep}}$) is derived. Specifically, draws are taken from the posterior predictive distribution tied to the replicated data, $\mathrm{D}^{\mathrm{rep}}$, in order to compute the probability of events that involve $\mathrm{D}^{\mathrm{rep}}$: $p(\mathrm{D}^{\mathrm{rep}}|\mathrm{M},\mathrm{D}) = \int p(\mathrm{D}^{\mathrm{rep}}|\mathrm{M},\theta)p(\theta|\mathrm{M},\mathrm{D})d\theta,$ where θ is the unknown model parameters. This equation serves as the reference distribution by which one may evaluate the model. Having formulated a posterior distribution of θ based on the model and data, a predictive distribution of $\mathrm{D}^{\mathrm{rep}}$ can be used to "check" the model.

In PPMC, many data sets are replicated according to the same model M and a sample of possible θ values that have been drawn the posterior distribution. The resulting $\mathrm{D}^{\mathrm{rep}}$ replications form the posterior predictive reference distribution, which can be used to explore the usefulness of a model in analyses of future data that are somewhat similar to the observed data. In that sense, the Bayesian concept of PPMC is closely related to the parametric bootstrapping procedure described earlier, though the reference distribution for data replication is the posterior predictive distribution rather than the sampling distribution.

Test statistics and test quantities. In PPMC, failings of the model are indicated by the presence of systematic differences between the observed and replicated data. To assess such differences, Bayesian

researchers select statistical measures that represent certain aspects of the data and/or model that are deemed relevant to the topic under investigation. Borrowing the nomenclature of Gelman et al. (2013), such measures are referred to as *test statistics* (denoted by $T(\mathrm{D})$) if they depend on the observed data alone. For example, PPMC could be used to examine whether the median (or extreme cases, quantiles, or any other data-based test statistic) of the observed data is likely, given the reference distribution. When model checking measures depend on both model and data, they are referred to as *test quantities* and denoted by $T(\mathrm{M},\mathrm{D})$. For example, GOF indices by definition require a model and some data. When PPMC is applied to a test quantity, it yields $\mathrm{D}^{\mathrm{obs}}$: a simulated sample of values obtained by fitting the model to the observed data with different sets of parameter values drawn from the posterior distribution.

Bayesian model evaluation proceeds by comparing an observed test statistic or quantity T to the same test statistic or quantity from the reference distribution of $\mathrm{D}^{\mathrm{rep}}$. A posterior predictive *p*-value (or *ppp*-value) is then computed to quantify the likelihood of T in the reference distribution: $ppp = p\big(T(\mathrm{D}^{\mathrm{rep}}) \geq T(\mathrm{D})\big)$ or $ppp = p\left(T(\mathrm{D}^{\mathrm{rep}}) \geq T\big(\mathrm{D}^{\mathrm{obs}}\big)\right)$. The *ppp*-values denote the similarity between the observed and replicated data, relative to the chosen test statistic or quantity. The presence of systematic differences between the observed and predictive values is indicated by $ppp \leq .05$ or $\geq .95$; *ppp*-values near .50 indicate that there are no differences between the observed T and the distribution of T values that one would expect if the model were correct (Stone & Zhu, 2015). Although these cutoff values are commonly implemented, we do not advocate the strict use of cutoff values in general. The main goal here is to evaluate the degree to which a model minimizes the number of extreme *ppp*-values.

Finally, the analysis below illustrates an application of PPMC to evaluate various statistical aspects of a single model, but PPMC can also be used for model comparison (e.g., Béguin & Glas,

2001; Sinharay, 2006). In that case, one could inspect the relative fit of competing models by examining the frequency of extreme *ppp*-values in each model; the model with fewer *ppp*-values below .05 or above .95 would be deemed to better represent the observed data, with regard to the particular T under investigation. For further details and an example of PPMC-based model comparison in the context of IRT modeling, see Zhu and Stone (2012).

Methods

Data. As an illustration, we used secondary data[2] from the Sequenced Treatment Alternative to Relieve Depression (STAR*D; Rush et al.., 2004) trial, which was supported by the National Institute of Mental Health. Data were collected from 41 clinical outpatient patient facilities in the United States. Participants in the STAR*D trial completed the 139-item self-report Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001), which uses DSM-IV criteria to measure 15 psychiatric disorders, including major depressive disorder (MDD). Further details can be found at https://www.nimh.nih.gov/funding/clinical-research/practical/stard/index.shtml.

We analyzed the dichotomous (*Yes*/*No*) responses to the MDD items from a random sample of $N = 1000$ STAR*D participants. For the purposes of our illustration, three MDD items were removed from the original PDSQ data. Item 1 ("During the past two weeks, did you feel sad or depressed?") had an extremely high endorsement rate of 98.22% and was therefore deemed uninformative. Further, as in Bonifay (2015), Items 6 ("Was your appetite significantly greater than usual nearly every day of the past 2 weeks?") and 8 ("Did you sleep at least 1-2 hours more than usual nearly every day of the past 2 weeks?") were also removed; these items were worded almost identically to Items 5 and 7 (about "lower" appetite and "less" sleep, respectively) and thus highly redundant.

Model. Practitioners often rely on PDSQ raw score thresholds for potential diagnosis. For example, Castel et al. (2007) examined the specificity and sensitivity of several PDSQ subscales in order

[2] Thank you to Dr. Waguih IsHak of the Geffen School of Medicine at UCLA for providing this data set.

to ascertain raw cut-off scores for clinical diagnosis and were able to approximately replicate the MDD threshold that is recommended in the PDSQ manual. Importantly for our illustration, this use of raw scores implies that the MDD item set is unidimensional (McNeish & Wolf, 2020) in that all items contribute to the measurement of a single factor (i.e., "depression"). The assumed unidimensionality of the MDD subscale is also evident in several studies involving the MDD subscale itself (e.g., Houben et al., 2017) as well as research on the factor structure of the full PDSQ wherein all MDD items were measured by one factor (e.g., Gibbons, Rush, & Immekus, 2009). We followed this usual approach by specifying an item response theory (IRT) model with a unidimensional test-level structure.

At the item-level, a variety of IRT models can be applied to estimate the probability of MDD symptom presence (or absence). We considered the two-parameter logistic model (2PLM; Birnbaum, 1968), which is among the simplest (and most widely applied) IRT models. The 2PLM is given by

$$P(x_{ij} = 1|\vartheta_j; a_i, b_i) = \frac{\exp[a_i(\vartheta_j - b_i)]}{1 + \exp[a_i(\vartheta_j - b_i)]}, \quad (2)$$

where the probability P of a correct response $x = 1$ on item i by person j is conditional on the location ϑ of person j along the latent trait continuum and two item parameters. The a_i (or "discrimination") parameter represents the degree to which item i differentiates between respondents at lower and higher latent trait locations. The b_i (or "difficulty") parameter represents the location of the item along the latent trait scale.

In sum, the model under evaluation in this paper specifies a 2PLM for each item, a unidimensional structure for the complete set of items, and a normally distributed latent variable underlying the response patterns. For ease of communicating our findings, we will use the shorthand "Mod.U2N" to denote a unidimensional 2PLM with a normally distributed person parameter.

Estimation. To compare the model checking approach against more traditional model evaluation criteria, we implemented two estimation strategies. First, we used the default specification in

flexMIRT: marginal maximum likelihood estimation via the expectation-maximization algorithm (Bock & Aitkin, 1981) with 49 quadrature points equally spaced between values of -6.0 and 6.0 on the latent trait scale. Second, we conducted Bayesian estimation via Metropolis-within-Gibbs Markov chain Monte Carlo simulation. For simplicity, we accepted the default prior distributions: flat (non-informative) priors for the item parameters and a standard normal prior for the person parameter (i.e., $\vartheta \sim N(0,1)$).[3] Reliance on default prior distributions is generally ill-advised (Kass & Wasserman, 1996), and given the characteristics of the STAR*D participants (i.e., clinical outpatients), a normally distributed prior may be a model misspecification. However, (a) a normal prior on the person parameter is the default in standard IRT person parameter estimation (i.e., EAP or MAP "scoring") and thus likely to be applied in practice, and (b) PPMC allows one to empirically investigate whether this prior is likely to be correct or not (Gelman & Shalizi, 2013).

Traditional model evaluation. For the sake of comparison, we calculated traditional goodness-of-fit statistics by fitting Mod.U2N to the observed data via Metropolis-Hastings Robbins-Monro estimation (Cai, 2010) in flexMIRT. Note that the 18 dichotomous MDD items permitted $2^{18} = 262{,}144$ possible response patterns (i.e., an extremely sparse contingency table), thus precluding consideration of full-information GOF statistics. Accordingly, we consulted the ordinal M_2* statistic (Cai & Hansen, 2013; Maydeu-Olivares, 2013), which is asymptotically χ^2 distributed with $\kappa - \nu$ degrees of freedom (df), where κ is the number of reduced first- and second-order marginal residuals (i.e., $\kappa = n(n+1)/2$ for n items) and ν denotes the number of free parameters. We also considered the root mean square error of approximation based on the M_2^* statistic ($RMSEA_2$), which is given by $\sqrt{(M_2^* - df)/(N \cdot df)}$, where N is the sample size. Conveniently, $RMSEA_2$ is interpreted just as the

[3] One important aspect when deciding on the specification of priors is the prior-data disagreement issue that can arise. If priors are misaligned with the evidence in the data, then the posteriors can be impacted by the priors, as well as GOF. This issue is particularly common if informative, but "inaccurate," priors are implemented. We used non-informative priors to avoid this issue. For more on data-prior conflict, please see Evans and Moshonov (2006).

traditional RMSEA: values below .08 represent mediocre fit and values below.05 support good fit (MacCallum, Browne, & Sugawara, 1996).

Bayesian model evaluation. PPMC can certainly be used in an omnibus sense, by evaluating the general GOF of the whole model; however, as Gelman et al. (1996) noted, "Indeed, Bayesian inference is a powerful tool for learning about model defects, because we have the ability to examine, as a discrepancy measure, any function of data and parameters" (p. 758). The potential to examine any feature of the data highlights the versatility of the model checking method: Test statistics and quantities can be chosen by the researcher in order to assess some characteristic of the data that is not directly addressed by the probability model. For example, rather than considering the mean or variance, one may wish to explore the degree to which the rank ordering of the observed sample is the same in the replicated distribution as in the observed data. Similarly, it may be informative to know whether the residuals from fitting the model to the observed data are similar in size to those obtained from the replicated data.

To demonstrate PPMC, we selected two test statistics (based only on the observed data) and two test quantities (based on the data and model). Note that the choice of test statistics or quantities is up to the researcher, and any feature of the data and/or model can be evaluated via PPMC. Though we recommend that feature selection should have some theoretical justification. For example, if an educational assessment includes open-ended constructed response items (meaning correct answers cannot be attributed to guessing), then it would be uninformative to use PPMC to investigate whether the lower asymptote (or "guessing") parameter of a dichotomous IRT model is greater than zero.

Test statistics. In the present example, we decided to first examine the item response proportions, which are simply the proportions of *Yes* endorsements out of all valid responses to each MDD item. Mod.U2N specified a symmetric (normal) prior distribution for the person parameter (i.e., the latent variable of "depression"). To be clear, analysis of the item response proportions will indicate

symmetry but will not provide evidence regarding the normality of the latent trait (because, e.g., a bimodal distribution may be symmetric but not normal). Though the flexibility of PPMC would certainly allow one to examine this issue more formally, by checking the model according to recent advancements in latent variable distribution fit (e.g., Monroe, in press; Zhen & Cai, 2018). More simply, it may be the case that the levels of depression among respondents in the observed data are not symmetrically distributed. Perhaps the STAR*D sample was characterized by many patients with higher levels of depression (more *Yes* responses) and few with low levels (fewer *No* responses) than one would expect in a symmetric distribution. In such a case, the observed response proportions would be higher than those that were generated from the posterior predictive distribution D^{rep} given Mod.U2N, thus suggesting that a skewed prior may be a more appropriate specification. To investigate this possibility, we conducted PPMC of the item response proportions, comparing the observed response proportions in the data to the replicated response proportions in D^{rep}.

Our second test statistic was the item-total correlation (i.e., correlation between each MDD item and the total score for all items). While the response proportions enabled us to check the person parameter, the item-total correlations allow us to check an important item property: If all items are measuring a single construct, then they should be similarly correlated with the total scores. It is trivial to establish that this is not the case in the observed data, with item-total correlations ranging .30 to .59. However, PPMC of the item-total correlations goes beyond analysis of the observed data by presenting a distribution of item-total correlations that we would expect to see if Mod.U2N were correct.

Test quantities. IRT analyses provide several pieces of output that qualify as suitable test quantities. We first checked the 2PLM by considering the $S\text{-}X_i^2$ item-fit statistics (Orlando & Thissen, 2000), which indicate the degree to which the item response function is under- or overestimating the empirical proportion of endorsements of item i. For PPMC of the fit of the 2PLM to each item, we calculated how often the $S\text{-}X_i^2$ values in D^{rep} exceeded the $S\text{-}X_i^2$ values in D^{obs}. Support for the 2PLM

as an appropriate item-level model was indicated whenever the observed data yielded item-fit statistics that were equivalent to the data replicated under Mod.U2N (i.e., items with *ppp*-values near .5).

Finally, we also considered the degree to which Mod.U2N was able to capture the residual correlations (*aka* local dependence (LD)) between each MDD item pair. By definition, a unidimensional IRT model is not equipped to capture any meaningful LD among the items. To check this in Mod.U2N, we examined the LD X^2 index (Chen & Thissen, 1997), which is based on the ϕ correlation matrices of the observed and model-implied bivariate contingency tables. When the observed correlation is higher than the model-implied correlation for an item pair, the result is positive LD; if the model-implied correlation is higher, then negative LD has been detected within that item pair. For PPMC, *ppp*-values were obtained for all LD X^2 values in every data set in D^{obs} and D^{rep}.

Analytic strategy. Model evaluation via PPMC can be performed using any Bayesian statistical software. Because our aim was to evaluate an IRT model, we conducted all simulations and analyses in flexMIRT v3.6 (Cai, 2020), which we called from within R (R Core Team, 2017) via the irtplay package (Lim & Wells, 2020). The R script comprised six steps. First, we performed a Bayesian analysis of the MDD data according to the model specifications described below, and saved the output from each of 1,000 Markov chain Monte Carlo (MCMC) iterations. Second, we used the parameter vector from each MCMC cycle to generate 1,000 unique data sets. In the third and fourth steps, we obtained test statistics/quantities from the observed and replicated data, respectively. Fifth, we calculated the *ppp*-values and associated descriptives of several classical and IRT item-level and test-level statistics. Finally, we plotted the key results. The complete annotated R script is available at https://osf.io/42cz7/.

Results

Due to space constraints and because they were not our primary concern, we have opted not to present the parameter estimates that we obtained through either traditional or Bayesian estimation. In the traditional analysis, all a_i parameters (M = 1.55, SD = 1.30) were significantly different than

zero and b_i parameters ranged from -3.56 to 1.42. Bayesian estimates of the a_i (M = 1.58, SD = 1.34) and b_i (range: -3.52, 1.49) parameters were comparable to the traditional estimates. In practice, these results should be closely inspected, but for this illustration, we are bypassing the parameter estimates so that we can emphasize model evaluation.

Traditional model evaluation. According to traditional GOF assessment, the MDD items were not well represented by Mod.U2N, $M_2(135) = 1192.88$, $p < .0001$, $RMSEA_2 = .09$. Though somewhat informative, these results do not provide much insight into the failings of Mod.U2N; they tell us that the overall fit of Mod.U2N to the particular data in the STAR*D sample is weak, but they provide no details regarding the precise weaknesses of this model. Of course, we could continue investigating the data from a traditional perspective, examining item-level statistics, residuals, and other IRT output, but such results inform us only about the observed data. A deeper understanding of Mod.U2N, that is, beyond GOF and without strict reliance on the observed data, can be gleaned from PPMC.

Bayesian model evaluation: Test statistics. Table 1 displays the observed response proportions for each of the 18 MDD items (paraphrased for simpler presentation), along with the PPMC results. All proportions in the observed data fell within the uncertainty interval of the posterior distribution and the corresponding *ppp*-values (ranging from .417 to .673) indicated that the replicated proportions were fairly indistinguishable from the observed proportions. In other words, Mod.U2N is able to generate many data sets with response proportions that are highly similar to those that we observed. We can thereby conclude that the observed response proportions are not driving the poor fit.

Table 2 presents the observed and replicated item-total correlations. Unlike the response proportions, the item-total correlations showed that the observed sample yielded data with unexpectedly high (Items 1, 2, 3, 11, and 12) or low (Items 13-18) correlations with the total MDD scores. For example, Item 2 ("Did you get less joy or pleasure from almost all of the things you normally enjoy?") had a higher item-total correlation in the observed data (r = .408) than in all but 11 of the 1,000

replicated data sets (i.e., *ppp* = .011). Conversely, the item-total correlations of Items 14 ("Did you wish you were dead?") and 15 ("Did you think you'd be better off dead") were higher (r = .630 and .639, respectively) than observed in 100% of the replicated data sets (i.e., *ppp* = 1.000). Overall, these results indicate that data consistent with Mod.U2N would have had substantially different item-total correlations than those obtained from the STAR*D participant data. Therefore, PPMC of the item-total correlations provides evidence that Mod.U2N is not likely to have generated the observed data.

Bayesian model evaluation: Test quantities. It is a bit more complicated to convey PPMC findings in the context of test quantities. The results above were found by comparing a test statistic from a single observed data set to the distribution of that same statistic in many replicated data sets; however, model-based test quantities such as item-fit or local dependence statistics require an inspection of the complete D^{obs} distribution. Then, PPMC proceeds by simply counting the number of replications in which the test quantity obtained via D^{rep} exceeds that of D^{obs}. Consequently, PPMC of test quantities can be communicated by comparing the distributions of D^{obs} and D^{rep} and/or by focusing primarily on *ppp*-values.

The results from PPMC of the $S\text{-}X_i^2$ item-fit index are displayed in Figure 2. Ideally, each item would resemble Item 3 ("Did you get less joy or pleasure from almost all of the things you normally enjoy?"), in which the $S\text{-}X_i^2$ density plots for D^{obs} (dark fill) and D^{rep} (light fill) were almost perfectly superimposed (*ppp* = .512). However, we found that the observed item-fit of Items 13-16 was substantially higher that we would anticipate if Mod.U2N were correct. The expected misfit of Item 15 ("Did you think you'd be better off dead?") was particularly egregious: In D^{obs}, the expected $S\text{-}X_i^2$ was 122.359 (95% UI = 76.268, 197.55), while in D^{rep}, the expected $S\text{-}X_i^2$ was just 10.363 (95% UI = 3.198 to 22.805), and in fact, there was not a single replication in which Item 15 yielded worse item-fit in D^{rep} than in D^{obs}. Further, though the D^{obs} values were technically within the D^{rep} uncertainty intervals, the *ppp*-values of Items 2, 4, 17, and 18 were also far smaller than expected under Mod.U2N. Thus,

PPMC of the $S\text{-}X_i^2$ quantities provides compelling evidence against the 2PLM as an appropriate item-level model for several of the MDD items.

Finally, Figure 3 illustrates the *ppp*-values of the local dependencies between all $18(18 - 1)/2 = 153$ item pairs. In this figure, pie charts that are close to a 50/50 split and darker in color (e.g., the chart for Items 3 and 4) reflect LD X^2 quantities that were similar for D^{obs} and D^{rep}, meaning that the observed LD of that particular item pair would be likely to replicate if Mod.U2N were correct. Recall that Mod.U2N was specified as a unidimensional model, such that the responses to all MDD items were hypothesized as reflecting a single latent continuum (e.g., "depression"). If that were correct, then all pies in Figure 3 would be evenly halved. Clearly, that is not the case: Many pairs are depicted as slivers, lines, or even dots, indicating that the LD X^2 of D^{rep} was rarely, almost never, or never greater than or equal to that of D^{obs}, respectively. In summary, PPMC of local dependence revealed an additional flaw of Mod.U2N: A single latent dimension is unable to account for the residual correlations that were evident in the observed data.

Discussion

The aim of this paper was to showcase the utility of the Bayesian model evaluation technique known as posterior predictive model checking, and to illustrate the insights that PPMC offers in the context of categorical data using an example pertinent to prevention science. Rather than relying on traditional goodness-of-fit assessment, which is a thorny issue in the presence of categorical data and is fraught with methodological problems, we used PPMC to explore statistical issues related to a dichotomous measurement model of 18 major depressive disorder items from the Psychiatric Diagnostic Screening Questionnaire. PPMC allows users to scrutinize any aspect of their data or model, and we demonstrated this technique by examining two data-based test statistics (response proportions, item-total correlations) and two IRT model-based test quantities (item-fit, LD). The PPMC results have implications for specific methodological issues and prevention science research more broadly.

Methodological implications. The observed item response proportions (Table 1) were aligned with what we would expect given our hypothesized model: a unidimensional 2PL IRT model with a normally distributed person parameter. However, the PPMC results identified several problematic items, both in terms of unexpected item-total correlations (Table 2) and poor representation by the 2PLM (Figure 2), as well as overall misspecification regarding the underlying dimensionality of the item set (Figure 3). This latter finding is especially revealing, as Gelman and Shalizi (2013) argue that the subjectivity of the Bayesian prior (a deservedly major sticking point for many traditional practitioners) can be investigated via PPMC of the prior itself. In our study, the results provided evidence that a univariate normal prior of the person parameter was highly unlikely to produce the observed response patterns. Further research would be needed to determine whether the prior was errant with regard to unidimensionality, normality, or their interaction. Although the STAR*D participants comprised a clinical outpatient sample, we would argue that individual differences in the degree of MDD could still be normally distributed. Although the *ppp*-values of the item response proportions (Table 1) do not directly assess the normality of the latent trait, we did find that the proportions in data sets that were replicated under assumed normality were virtually indistinguishable from the observed proportions. If the person parameter had been asymmetric, for example negatively skewed such that the sample included more people with high levels of depression, then the posterior predictive replications would have yielded lower response proportions than in the observed data. However, none of the *ppp*-values in Table 1 suggest that this was the case; whether this symmetry reflects a normally distributed latent variable is a topic for future research.

It may be that the unidimensionality specification is the culprit (among the four data/model features that we checked; further PPMC may reveal additional shortcomings of Mod.U2N). Importantly, if the MDD item set is multidimensional, then several psychometric methods that assume unidimensionality would be invalidated, including Cronbach's alpha (McNeish, 2018) and commonly

applied measurement techniques such as linking and equating (Bolt, 1999) and computerized adaptive testing (Ackerman, 1991). In the specific context of IRT, the precision and accuracy of item parameter estimates (Way, Ansley, & Forsyth, 1988) and person parameter estimates/factor scores (Ansley & Forsyth, 1985) are compromised when multidimensional data are represented by a unidimensional model. Further adverse consequences of ignoring multidimensionality in item response data can be found in Reise, Cook, and Moore (2015).

An issue that is especially germane to our MDD illustration was recently illuminated by McNeish and Wolf (2020): Summed scores, such as those used in diagnosing psychopathology with the PDSQ, correspond to a constrained (parallel) latent variable model. From this correspondence between summed scores and factor scores, it follows that a violation of unidimensionality, such as that shown in Figure 3, would also violate the application of summed scores. This is particularly concerning in the context of the PDSQ, wherein diagnoses are based on simply tallying the responses to the symptom checklist. If the MDD item set is multidimensional, then the self-reported presence of a certain number of symptoms should not be used to inform clinical diagnosis. In our illustration, PPMC revealed that Mod.U2N does not support the degree of multidimensionality that is present in the observed data, and having uncovered this flaw in the model, we can recommend that summed scores (and other methods that assume a single underlying dimension) are not warranted or defensible.

Prevention science implications. Beyond the methodological considerations of the MDD items of the PDSQ, PPMC offers great utility in the science of prevention. Whenever a statistical model is used to represent a prevention science theory, it should be rigorously evaluated. As discussed, traditional GOF assessment and bootstrapping methods can be misleading due to their overreliance on the observed data. For example, if the aim is to prevent the development of depression, then the model of depression should be thoroughly evaluated, not only in terms of global fit to the data, but with regard to specific flaws. As Box (1976) noted (in a subsection entitled "Selective Worrying"): "Since all

models are wrong, the scientist must be alert to what is importantly wrong" (p. 792); PPMC allows users to examine precisely why the model is wrong, and to thereby improve it. In the prevention sciences, if we can better understand why a model is wrong, then we will be better equipped to improve our efforts at prevention.

Limitations and Future Directions. The main point of concern that some researchers have surrounding Bayesian inference is the subjectivity of specifying the prior. Bayes' theory does not place any constraints on how one should set the prior, so the choice of prior may differ from person to person, depending on each person's previous beliefs about the phenomenon under investigation. Consequently, the posterior, which is a function of the prior, can also differ from one researcher to the next. However, we note that subjectivity is not solely placed within the Bayesian estimation framework. It is also the case that the model being estimated is subjective in that it can be specified in a variety of ways, depending on the subjective opinions of the researcher. Thus, the model checking techniques discussed above are based on the subjective opinions that characterize the model, the priors, and the posterior distributions, which from the traditional perspective, may be uncomfortable for some researchers. However, it is our view that the advantages of PPMC outweigh the complications, especially in light of the potential for PPMC to directly test the appropriateness of the prior specification (Gelman & Shalizi, 2013). We also note that embracing subjectivity in data analysis can provide a rich set of results and a deeper understanding of the phenomena under investigation.

To that end, a specific limitation of this study is the use of the default (normal) prior. In application, it is a good strategy to assess the impact that prior settings have on results, especially in the case of subjective priors. If modifying the priors impacts the posteriors or model checking phase in a substantive way, then this would be an important aspect to note. One way of addressing this issue is to conduct a sensitivity analysis of the prior settings. A sensitivity analysis can be used to examine how robust posterior results are when the prior settings are altered. If posterior findings remain

relatively stable under statistically and substantively different prior settings, then there can be confidence that the impact of prior subjectivity is minimal. However, if findings vary when priors are altered, then the subjectivity of opinions plays a stronger role in the formation of the posterior, and potentially on model evaluation. We also note that a similar sensitivity analysis can be conducted on the likelihood, where the robustness of substantive results can be examined by performing a sensitivity analysis on the model—this process need not be confined to the Bayesian framework. For more information on these issues, see Depaoli and van de Schoot (2017).

As a final note, we would like to briefly address another common criticism of Bayesian methods: MCMC estimation and PPMC are not computationally efficient. The Bayesian analyses presented in this paper required far longer than traditional methods. The test quantities in particular were a considerable computation burden: To obtain D^{obs} and D^{rep} required fitting 1,000 IRT models, and even though these were relatively simple models, the amount of computation time was orders of magnitude longer than with the traditional approach (i.e., ~40 minutes vs. a few seconds). Despite this, we believe that the Bayesian approach is worth the effort. As discussed throughout this Special Issue, categorical data introduces several statistical and analytical complications, and proper evaluation of categorical data models is crucial, especially when the goal is as important as prevention. Thus, while Bayesian methods are undeniably more laborious than traditional model evaluation strategies, it is our view that the extra work pays off in more informative, revealing, and useable results.

Our primary motivation was to demonstrate the insights that are afforded by Bayesian model evaluation. Having conducted PPMC and uncovered various problems with the MDD item response data, an obvious future direction would be to apply a multidimensional model. Model evaluation becomes much more cumbersome in such an analysis, as each dimension must be appraised, scored, and interpreted. Fortunately, Bayesian methods are available for multidimensional IRT (Fox, 2010) and PPMC proceeds conceptually just as described herein (e.g., Levy, 2011). Thus, while it was not our

intention to uncover the true structure of the MDD items, future researchers could certainly build on our PPMC results to better understand the dimensionality (and thereby enhance the use and interpretation of scores) of the MDD items and to extend our results to the PDSQ as a whole.

*Conclusion***.** In summary, Bayesian methods are rarely applied in prevention science, but they offer several advantages. This paper focused on the use of Bayesian model checking to more thoroughly evaluate models of categorical data, which are notoriously difficult to appraise using traditional methods. Our findings demonstrate that PPMC can be an especially enlightening model evaluation strategy, and we suggest that this technique be added to the statistical toolbox of prevention scientists.

Compliance with Ethical Standards

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210032. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Ethics approval: All procedures performed in the STAR*D trial was approved by the institutional review board of the STAR*D National Coordinating Center at the University of Texas Southwestern Medical Center in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Conflicts: The authors declare no conflicts or competing interests.

Consent to participate: Informed consent was obtained from all participants.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, *15*(1), 13-24.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37-48.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research, 27*(4), 525-546.

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: AddisonWesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in education*, *12*(4), 383-407.

Bonifay, W. (2015). An illustration of the two-tier item factor analysis model. In S. P. Reise and D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 207–225). New York: Routledge.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research, 52*(4), 465-484.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335.

Cai, L. (2020). flexMIRT R version 3.6: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, Chung, and Lee (*in press*). Incremental model fit assessment in the case of categorical data: Tucker-Lewis Index for item response theory. *Prevention Science.*

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245-276.

Castel, S., Rush, B., Kennedy, S., Fulton, K., & Toneatto, T. (2007). Screening for mental health problems among patients with substance use disorders: Preliminary findings on the validation of a self-assessment instrument. *The Canadian Journal of Psychiatry*, *52*(1), 22-27.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Depaoli, S., Yang, Y., & Felt, J. (2017). Using Bayesian statistics to model uncertainty in mixture models: A sensitivity analysis of priors. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 198-215.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914. https://doi.org/10.1214/06-BA129

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer Science & Business Media.

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

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733-760.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8-38.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research, 43*, 401-410.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological),* 83-100.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing!

testing! one, two, three–Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841-850.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods* (Vol. 580). New York: Springer.

Houben, M., Claes, L., Vansteelandt, K., Berens, A., Sleuwaegen, E., & Kuppens, P. (2017). The emotion regulation function of nonsuicidal self-injury: A momentary assessment study in inpatients with borderline personality disorder features. *Journal of Abnormal Psychology, 126*(1), 89-95.

Kadane, J. B. (2015). Bayesian methods for prevention research. *Prevention Science, 16*(7), 1017-1025.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York: Guilford Press.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association*, *91*(435), 1343-1370.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24*(4), 492-516.

Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of Educational and Behavioral Statistics*, *36*(5), 672-694.

Lim, H., & Wells, C. S. (2020). irtplay: An R package for online item calibration, scoring, evaluation of model fit, and useful functions for unidimensional IRT. *Applied Psychological Measurement*, doi: 0146621620921247.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130-149.

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity, 28*(2), 185-217.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11*, 71-101.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412-433.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods*, 1-19.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Ory, D. T., & Mokhtarian, P. L. (2010). The impact of non-normality, sample size and estimation

technique on goodness-of-fit measures in structural equation modeling: Evidence from ten empirical models of travel behavior. *Quality & Quantity, 44*(3), 427-445.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. R., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 13-40). New York, NY: Routledge.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics, 9*(1), 130-134.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151-1172.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., . . . & Niederehe, G. (2004). Sequenced treatment alternatives to relieve depression (STAR* D): Rationale and design. *Controlled Clinical Trials, 25*(1), 119–142.

Stone, C. A., & Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models Using SAS®*. Cary, NC: SAS Institute Inc.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842-860.

van Erp, S., Mulder, J., & Oberski, D (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods, 23,* 363-388.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, *12*(3), 239-252.

Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire. *Archives of general psychiatry*, *58*(8), 787-794.