Research Article

Statistics in Medicine

Received 15 January 2010, Accepted 13 September 2010 Published online 6 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4109

Using latent variable modeling and multiple imputation to calibrate rater bias in diagnosis assessment

Juned Siddique,[a*†] Catherine M. Crespi,[b] Robert D. Gibbons[c] and Bonnie L. Green[d]

We present an approach that uses latent variable modeling and multiple imputation to correct rater bias when one group of raters tends to be more lenient in assigning a diagnosis than another. Our method assumes that there exists an unobserved moderate category of patient who is assigned a positive diagnosis by one type of rater and a negative diagnosis by the other type. We present a Bayesian random effects censored ordinal probit model that allows us to calibrate the diagnoses across rater types by identifying and multiply imputing 'case' or 'non-case' status for patients in the moderate category. A Markov chain Monte Carlo algorithm is presented to estimate the posterior distribution of the model parameters and generate multiple imputations. Our method enables the calibrated diagnosis variable to be used in subsequent analyses while also preserving uncertainty in true diagnosis. We apply our model to diagnoses of posttraumatic stress disorder (PTSD) from a depression study where nurse practitioners were twice as likely as clinical psychologists to diagnose PTSD despite the fact that participants were randomly assigned to either a nurse or a psychologist. Our model appears to balance PTSD rates across raters, provides a good fit to the data, and preserves between-rater variability. After calibrating the diagnoses of PTSD across rater types, we perform an analysis looking at the effects of comorbid PTSD on changes in depression scores over time. Results are compared with an analysis that uses the original diagnoses and show that calibrating the PTSD diagnoses can yield different inferences.

Keywords: Markov chain Monte Carlo; ordinal probit; PTSD; censoring; leniency error

1. Introduction

Whenever participants in a study are evaluated by different raters, the possibility exists that ratings may be biased—that is, raters may use different standards in applying the rating scale [1]. Beginning with Kneeland [2], researchers have been aware of the form of rater bias typically referred to as *leniency error*, the tendency for raters to be systematically lenient or harsh when making evaluations. Depending on the design of the study, leniency error can introduce bias into ratings, add additional variance to ratings, and inflate correlations across rating dimensions. In mental health research, leniency error can alter the prevalence of psychiatric diagnoses and can make comparisons among subjects difficult [3].

Several methods to correct leniency error have been proposed. Raymond and Viswesvaran [4] and de Gruijter [5] describe regression-based approaches that estimate predicted 'true' ratings while controlling for rater effects. Houston *et al.* [6] describe an imputation approach for imputing missing ratings so that every examinee is rated by every rater. An examinee's true score is then estimated by averaging over all raters—both imputed and observed. Braun [7] describes a design for calibrating rater bias that uses partially balanced incomplete block designs. All of these methods require that each subject is measured by more than one rater. However, in many studies, it is prohibitively expensive or infeasible for more than one rater to evaluate a subject. Furthermore, none of these methods correct ratings so that they can be used as a covariate in subsequent analyses.

[a]*Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, U.S.A.*
[b]*Department of Biostatistics, University of California Los Angeles School of Public Health, Los Angeles, CA, U.S.A.*
[c]*Center for Health Statistics, University of Illinois-Chicago, Chicago, IL, U.S.A.*
[d]*Department of Psychiatry, Georgetown University Medical Center, Washington, DC, U.S.A.*
[*]*Correspondence to: Juned Siddique, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N. Lake Shore Dr., Suite 1400, Chicago, IL 60611, U.S.A.*
[†]*E-mail: siddique@northwestern.edu*

We propose a method of correcting rater bias that can be used when each subject is assessed by a single rater, and allows the corrected diagnoses to be used in analyses while preserving uncertainty in true diagnosis. Our method uses a Bayesian random effects censored ordinal probit model to identify a latent moderate class of patients, who can be reclassified as cases or non-cases. A Markov chain Monte Carlo (MCMC) algorithm is presented to estimate the posterior distribution of the model parameters and generate multiple imputations for a recalibrated diagnosis variable. The recalibrated diagnosis variable can then be used in subsequent analyses. We apply our model to the diagnoses of posttraumatic stress disorder (PTSD) from a depression study in which nurse practitioners were twice as likely as clinical psychologists to diagnose PTSD despite the fact that participants were randomly assigned to either a nurse or a psychologist.

The outline for the remainder of this paper is as follows. In Section 2, we present our motivating example, the Women Entering Care (WECare) study. In the WECare study, diagnoses of PTSD were subject to rater bias due to raters having different clinical backgrounds. In Section 3, we present the latent variable framework that we will use to model diagnoses subject to rater bias and a Bayesian approach for estimating the model parameters and generating multiple imputations. In Section 4, we apply our methods to the WECare data to calibrate PTSD diagnoses across raters. We then reanalyze the WECare data with the calibrated diagnoses. In this reanalysis, PTSD is a moderator (covariate) in a model of depression scores over time. In Section 5, we summarize our conclusions and suggest additional areas of application. In the Appendix, we detail the MCMC algorithm used to estimate our model.

2. Motivating example: The WECare study

PTSD is an anxiety disorder that can develop after exposure to a terrifying event or ordeal in which grave physical harm occurred or was threatened [8]. Traumatic events that may trigger PTSD include violent personal/sexual assaults, natural or human-caused disasters, accidents, or military combat. PTSD affects about 7.7 million American adults [9], but it can occur at any age, including childhood [10]. Women are more likely to develop PTSD than men [9, 11], and there is some evidence that susceptibility to the disorder may run in families [12]. PTSD is often comorbid with depression, substance abuse, or one or more other anxiety disorders [13].

As is the case with diagnostic assessments more generally, assessments of PTSD are vulnerable to rater bias due to the fact that these assessments are, using the terminology of Hoyt and Kearns [1], *inferential*. That is, they utilize semi-structured interviews that require raters to make judgments about the meaning, representations, and severity of target emotions and behaviors. Hence, PTSD assessments allow more scope for disagreements among raters. As opposed to *explicit* rating systems which are based on counts or frequencies of readily observable behaviors, or checklists that simply require the respondent to reply 'yes' or 'no' to the presence of a symptom, inferential assessments of PTSD require the interviewer to obtain a detailed trauma history, and perform a fine-grained analysis of whether the respondent actually has the target symptom and not something else, and of symptom severity based on both the frequency and the intensity of the symptoms. This type of assessment draws heavily on the interviewer having a strong background in psychopathology, diagnosis, and structured interviewing, as well as extensive experience with the PTSD criteria specifically [14]. While explicit ratings systems for diagnosing PTSD do exist [15], they can only give a presumptive diagnosis. An inferential rating is required to produce a clinical diagnosis that is considered the gold standard.

Several studies have found that lay interviewers tend to overdiagnose PTSD when compared with diagnoses performed by expert clinicians. Breslau *et al.* [16] performed a comparison of PTSD assessed by lay interviewers using a structured interview with a diagnosis made by a clinician. They found that 25 per cent of cases diagnosed by lay interviewers as having PTSD did not have PTSD (as assessed by the clinician) and only 3 per cent of true PTSD cases were missed by the lay interviewers. They noted that the majority of the 'falsely' diagnosed cases were borderline PTSD cases, suggesting that the lay interviewers used a lower threshold than the clinicians.

The motivating example for this study is provided by the WECare study. The WECare Study was an NIH-funded trial that investigated depression outcomes during a 12-month period in which 267 low-income mostly minority women in the suburban Washington, D.C. area were treated for depression [17]. The participants were randomly assigned to one of three groups: medication, cognitive behavioral therapy (CBT), or treatment at usual (TAU), which consisted of a referral to a community provider. Participants randomized to medication were assigned to a nurse practitioner, participants randomized to CBT were assigned to a psychologist (Ph.D. in Clinical Psychology), and participants randomized to TAU were randomly assigned either to a nurse practitioner or to a psychologist. In each case, the clinician, prior to beginning treatment, conducted a clinical interview where PTSD was assessed using the PTSD module of the Structured Clinical Interview for DSM-III-R (SCID) [18]. Depression was measured every month for the first six months of the study and every month afterwards using the Hamilton Depression Rating Scale [19] through a phone interview by interviewers other than those performing treatment. The primary aim of the study was to investigate the effect of treatment on depression

Table I. WECare variables at baseline by treatment group.

Variable	Medication ($N=88$)	CBT ($N=90$)	TAU ($N=88$)	p-value for difference*
PTSD	60.2 per cent	34.4 per cent	48.9 per cent	0.002
Depression score	18.0	16.3	16.5	0.063
Anxiety score	16.2	15.1	14.3	0.173
Number of stressful life events	3.3	3.3	3.2	0.941
Rape	33.3 per cent	32.3 per cent	34.3 per cent	0.913
Age	28.7	29.8	29.5	0.607
Number of children	2.2	2.2	2.4	0.699
Has a partner	48.9 per cent	44.4 per cent	46.1 per cent	0.837
High school or more	58.0 per cent	70.0 per cent	60.7 per cent	0.217
Black	38.6 per cent	45.6 per cent	47.2 per cent	
Latina	54.6 per cent	47.8 per cent	48.3 per cent	0.766
White	6.8 per cent	6.7 per cent	4.5 per cent	

*Differences across continuous variables were tested using a one-way ANOVA. Differences across categorical variables were tested using a chi-squared test of independence. CBT: cognitive behavioral therapy; TAU: treatment as usual.

outcomes. A secondary aim was to examine whether the effects of treatment differed among women diagnosed with comorbid PTSD.

The developers of the SCID strongly suggest that it can only be used by professionals or clinically sophisticated paraprofessionals who have been highly trained in the use of the SCID [18]. In the WECare study, the investigators wanted the study participants to receive treatment from the same clinicians who performed the clinical interview. This required the nurse practitioners who provided the medication treatment to diagnose PTSD despite the fact that they did not have routine experience diagnosing psychiatric disorders or substantial experience with the SCID, although they did receive comprehensive training.

Table I displays the rates of PTSD and other baseline measures by treatment group. The medication group had almost twice the rate of PTSD as the CBT group (60 vs 34 per cent). The TAU group had a rate of comorbid PTSD that was about halfway between that of the Medication and CBT groups (49 per cent). Differences in other baseline measurements were not significant across the three groups at the 0.05 level.

The differences by treatment intervention assignment were perplexing because the clinical interview where PTSD was assessed took place prior to treatment. The reason behind the imbalance was revealed when PTSD rates were evaluated by the type of clinician (Table II). The PTSD rate was 60.6 per cent for nurse practitioners and 34.1 per cent for psychologists despite the fact that clinician assignment was randomized. Since approximately half the TAU clinical interviews were performed by nurses and half by psychotherapists, it is consistent that the TAU PTSD rate is close to the average PTSD rate of the two clinician types. The study investigators hypothesized that the reason for the difference in PTSD rates between the two types of clinicians was that the psychologists were using a higher threshold for PTSD. The psychologists in the study had worked in specialty care settings such as Veterans Administration hospitals, with patients with severe PTSD. The WECare participants assigned to psychologists were then judged based on these standards, although the WECare participants came from primary care settings where PTSD cases are less severe. As a result, the psychologists diagnosed PTSD at a low rate. The nurse practitioners on the other hand had less formal training in diagnosing PTSD, relying primarily on the training provided by the WECare investigators, and their limited experience with PTSD was in primary care settings with milder cases. For these reasons, the WECare investigators stated in their major outcomes paper [17] that the differential diagnosis rates between nurses and psychologists was due to leniency error. Specifically, they wrote:

> Despite comprehensive training on the standardized instrument (SCID), we found that nurse practitioners were significantly more likely to diagnose PTSD as compared with psychologists, suggesting that nurses used less stringent criteria in judging presence of symptoms.

An additional feature of the PTSD diagnoses not mentioned by the WECare investigators is that there is more between-rater variability among psychologists than among nurses. Among psychologists who diagnosed more than one subject, PTSD rates range between 10 and 63 per cent. Among nurses, PTSD rates range between 33 and 69 per cent. The between-rater standard deviation is 0.941 for psychologists, for nurses it is 0.626. This heterogeneity may be due to the fact that the psychologists were more likely to augment the training provided by the WECare investigators with their own subjective views regarding PTSD. Nurses, on the other hand, had little experience diagnosing PTSD and were more likely to rely on the training provided by the WECare investigators.

Green *et al.* [20] modeled depression outcomes in the WECare study using a random intercept and slope regression model where negative slopes indicate improvement in depression. Of particular interest to them were the three-way

Table II. PTSD rates by clinician.

Clinician	*N*	PTSD rate (per cent)	Between-rater standard deviation
Psychologist 1	11	45.5	
Psychologist 2	30	63.3	
Psychologist 3	20	10.0	
Psychologist 4	11	36.4	
Psychologist 5	35	17.1	
Psychologist 6	21	33.3	
Psychologist 7	1	100.0	
Psychologist subtotal	129	34.1	0.941
Nurse 1	6	33.3	
Nurse 2	52	63.5	
Nurse 3	49	69.4	
Nurse 4	13	53.8	
Nurse 5	17	41.2	
Nurse subtotal	137	60.6	0.626
Total	266	47.7	1.024

The overall rate of PTSD for psychologists is 34.1 per cent. For nurses it is 60.6 per cent. Psychologists exhibit more between-rater variability in their diagnoses than nurses: the between-rater standard deviation among psychologists is 0.941, for nurses it is 0.626.

treatment by month by PTSD interactions for medication and CBT. For each treatment, this term measures the difference in slopes between participants with and without PTSD. It was hypothesized that the presence of comorbid PTSD would moderate the effect of treatment on depression. Specifically, PTSD was thought to reduce the effectiveness of CBT on depression, resulting in a significant positive coefficient for the three-way CBT by month by PTSD interaction term, since the CBT therapy for the study was not designed to treat PTSD. However, since the medication used in the study, paroxetine, is FDA-approved for the treatment of both depression and PTSD, it was thought that comorbid PTSD would not have an influence on the effect of medication on depression, resulting in a coefficient for the three-way medication by month by PTSD interaction term not significantly different from zero.

Green *et al.* [20] attempted to correct for the imbalance in PTSD diagnoses by requiring both a positive diagnosis of PTSD on the clinical interview and a baseline Hamilton Anxiety Rating Scale score greater than 13 to classify a participant as having PTSD. This approach reduced the overall prevalence of PTSD, but a significant difference in PTSD rates between the two clinician types remained (psychologists = 24.0 per cent, nurses = 43.8 per cent, $p<0.001$). The analysis by Green *et al.* [20] resulted in three-way treatment by month by PTSD interactions for medication and CBT that were not significant. We sought to reevaluate the three-way interaction terms after calibrating the PTSD diagnoses between nurses and psychologists so that the rates were similar across these two groups.

3. Methods

3.1. Model

We assume that there exists an unobserved moderate diagnosis category between absence of the condition and presence of the condition. One type of rater, termed 'high-threshold raters', uses a high threshold and treats participants in this moderate category as non-cases, leading to low rates of diagnosis. Raters of the other type, 'low-threshold raters', have a lower threshold and treat participants in the moderate class as cases, leading to higher rates of diagnosis. By identifying the unobserved rater thresholds, using these thresholds to identify those patients who fall into the latent moderate class and reclassifying those patients as either cases or non-cases using other covariates in the data set, we can recalibrate the diagnosis rates between types of raters and correct the imbalance. For example, in the WECare study, to calibrate the nurses' rates to match those of the psychologists, we would identify the nurses' moderate diagnoses and reclassify them as non-cases.

Under our model, a positive diagnosis from a high-threshold rater is assumed to have a positive predictive value of 100 per cent; that is, a patient given a positive diagnosis by a high-threshold rater most certainly has the condition. The negative diagnoses of the high-threshold raters, on the other hand, are a mixture of non-cases and moderate cases. Conversely, since the low-threshold raters use much less stringent criteria, if they determine that a patient does not

have the condition then the patient almost certainly does not. Thus, we assume that the negative predictive value of a diagnosis from a low-threshold rater is 100 per cent. The positive diagnoses of the low-threshold raters are a mixture of severe cases and moderate cases.

Let y_{ijk} denote the true state of patient k diagnosed by rater j of rater type i, with 0=absence of condition, 1=moderate condition, and 2=severe condition. Let D_{ijk} denote the observed diagnosis (1=yes, 0=no) for patient k, rater j, rater type i. For both y_{ijk} and D_{ijk}, $k=1,\ldots,n_{ij}$, $j=1,\ldots,m_i$, $i=1$, if the rater is a low-threshold type and $i=2$ if the rater is a high-threshold type. Owing to the differential diagnosis thresholds, we have:

$$y_{ijk}=\begin{cases}0 & \text{if } D_{1jk}=0 \text{ (negative diagnosis by a low-threshold rater)}\\ 1 \text{ or } 2 & \text{if } D_{1jk}=1 \text{ (positive diagnosis by a low-threshold rater)}\\ 0 \text{ or } 1 & \text{if } D_{2jk}=0 \text{ (negative diagnosis by a high-threshold rater)}\\ 2 & \text{if } D_{2jk}=1 \text{ (positive diagnosis by a high-threshold rater)}\end{cases} \tag{1}$$

Thus, the observed positive diagnoses of the low-threshold raters are a mixture of moderate and severe cases and the negative observed diagnoses of the high-threshold raters are a mixture of non-cases and moderate cases.

A useful concept when modeling ordinal variables is that of an underlying latent variable [21]. Let z_{ijk} be a latent variable that measures symptoms associated with the condition on a continuous scale for participant k diagnosed by rater j of rater type i. Define

$$z_{ijk}=\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}+b_{j(i)}+\varepsilon_{ijk} \tag{2}$$

where $\boldsymbol{x}_{ijk}$ is a vector of explanatory variables (such as the variables other than PTSD in Table I), $\boldsymbol{\beta}$ is a regression parameter vector, ε_{ijk} is an error term, and $b_{j(i)}$ is a rater-specific random intercept where the subscript $j(i)$ indicates that the raters are nested within rater type (e.g. nurse, psychologist).

Let γ_1, γ_2, be cutoffs that determine which true category a participant falls into. Then

$$y_{ijk}=\begin{cases}0 & \text{(absence of condition)} \quad \text{if } z_{ijk}\leqslant\gamma_1\\ 1 & \text{(moderate condition)} \quad \text{if } \gamma_1<z_{ijk}\leqslant\gamma_2\\ 2 & \text{(severe condition)} \quad \text{if } z_{ijk}>\gamma_2\end{cases} \tag{3}$$

We assume in equation (2) that $\varepsilon_{ijk}\sim\mathrm{N}(0,1)$ and the random effects $b_{j(i)}$ are normally distributed with mean zero and different variance terms for different rater types. That is, $b_{j(i)}\sim\mathrm{N}(0,\sigma_i^2)$. The random rater effect $b_{j(i)}$ in equation (2) acts as a random threshold term so that each rater has their own threshold. This allows us to model between-rater variability such as that in Table II where the between-psychologist variance is different from the between-nurse variance. Equations (2) and (3) above describe a random effects censored ordinal probit model [22].

Figure 1 provides an illustration of the effect of differential thresholds on diagnosis rates. The three normal curves on the left side of the figure represent the distribution of the latent variable. Note that the mean and variance of the curve are the same in all three plots. The bar graphs on the right side of the figures are the rates of the diagnosis based on the location of the threshold. In the top row, the threshold (indicated by the dotted line) is to the right of the mean of the latent curve and the probability of a positive diagnosis is the area under the curve to the right of the threshold which is shaded black. As a result, the corresponding diagnosis rate is low. This is the mechanism that leads to low rates by high-threshold raters.

The middle row illustrates how placement of the threshold (again indicated by a dotted line) at a small latent value corresponds to high diagnosis rates. The bottom row plots both the high threshold and the low threshold from the plots above on the same latent curve. The area in black under the curve to the right of the high threshold is the probability of having a severe condition, the gray area under the curve to the left of the low threshold is absence of the condition. The lined area under the curve between the two thresholds is the probability of having a moderate condition. By estimating both thresholds, we can identify this moderate category condition.

Figure 1 also illustrates the censoring mechanism behind the diagnoses. Comparing the plots in the top and bottom rows, we see that the probability of a negative diagnosis by the high-threshold rater (top plot, gray area) consists of the sum of the probabilities of absence of condition and moderate condition (bottom plot, gray and lined areas). Comparing the middle and bottom rows, we see that the probability of a positive diagnosis by a low-threshold rater (middle plot, black area) consists of the sum of the probabilities of moderate condition and severe condition (bottom plot, lined and black areas).

Let $\mathscr{D}$ be the set of participants whose y_{ijk} values were observed, that is, they received a positive diagnosis from a high-threshold rater ($y_{2jk}=2$) or a negative diagnosis from a low-threshold rater ($y_{1jk}=0$). Let $\mathscr{R}$ (for right-censored) be the set of participants who received a positive diagnosis from a low-threshold rater and let $\mathscr{L}$ (for left-censored) be the

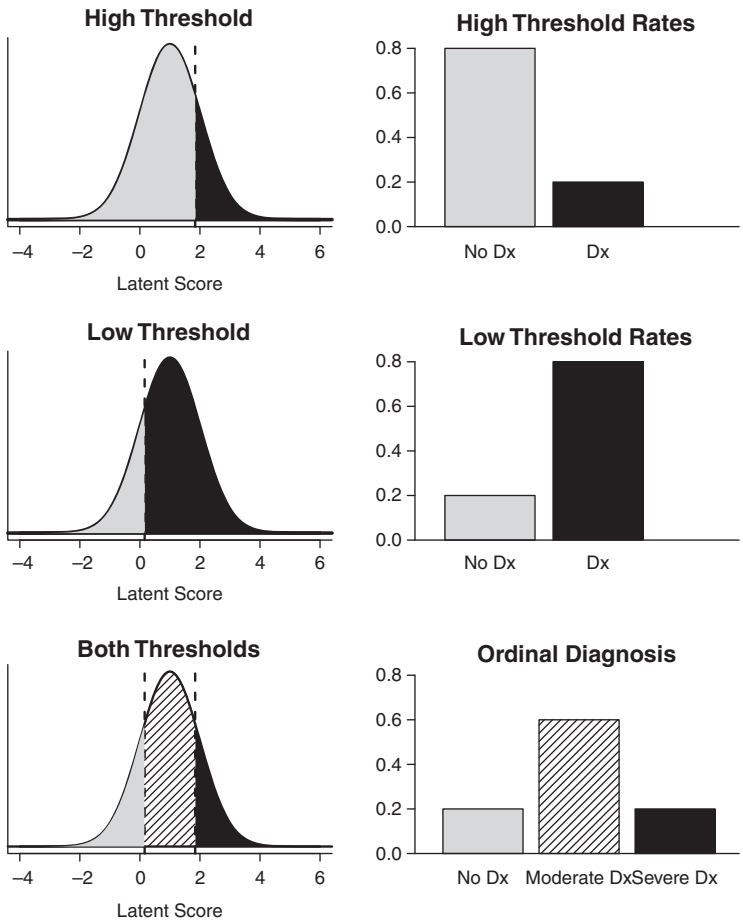

Figure 1. Illustration of the effect of differential thresholds on diagnosed rates. A high threshold leads to low diagnosed rates (gray in top row). A low threshold leads to high rates (black in middle row). By estimating both thresholds, we can identify a moderate category (lined in bottom row). Dx: Diagnosis.

set of participants who received a negative diagnosis from a high-threshold rater. Let $\boldsymbol{\gamma}$ be the vector of cutoff parameters (γ_1, γ_2) in equation (3), $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2)$ be the random effect variances for low- and high-threshold raters, respectively, and $\boldsymbol{z}$ be the vector of latent disease symptoms. The marginal likelihood of the latent data $\boldsymbol{z}$ along with the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ may be expressed as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}) = \prod_{i=1}^{2} \prod_{j=1}^{m_i} \int_{b_{j(i)}} \Bigg\{ & \prod_{k \in \mathscr{Q}} \phi(z_{ijk} - x_{ijk}^{\mathrm{T}} \beta - b_{j(i)} | b_{j(i)}) / (\gamma_{y_{ijk}} \leqslant z_{ijk} < \gamma_{y_{ijk}+1}) \\ & \times \prod_{k \in \mathscr{L}} \phi(z_{ijk} - x_{ijk}^{\mathrm{T}} \beta - b_{j(i)} | b_{j(i)}) / (z_{ijk} \leqslant \gamma_2) \\ & \times \prod_{k \in \mathscr{R}} \phi(z_{ijk} - x_{ijk}^{\mathrm{T}} \beta - b_{j(i)} | b_{j(i)}) / (z_{ijk} > \gamma_1) \Bigg\} f(b_{j(i)}) \mathrm{d} b_{j(i)} \end{aligned} \tag{4}$$

where $I(\cdot)$ indicates the indicator function, $\gamma_0 = -\infty$, $\gamma_3 = \infty$, ϕ denotes the standard normal density, and $f(b_{j(i)})$ is the density function of the random effect $b_{j(i)} \sim \mathrm{N}(0, \sigma_i^2)$.

3.2. Bayesian estimation

An ordinal regression model with three categories and two unknown cutoff parameters is overparameterized. Following Johnson and Albert [21], we resolve this identifiability problem by fixing the value of γ_1 at 0. The value of 0 is chosen for convenience. Let $\boldsymbol{b}_i$ be the vector of random effects $(b_{1(i)}, b_{2(i)}, \ldots, b_{m_i(i)})$ for rater type i and $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$. Assuming diffuse uniform priors taken on $\boldsymbol{\beta}$, γ_2, and σ_i, we can use a Gibbs sampling [23] approach for simulating from

the joint posterior of $(\boldsymbol{\beta}, \boldsymbol{b}, \gamma_2, \sigma^2, \boldsymbol{z}, \boldsymbol{y})$ using the following conditional distributions: $f(\boldsymbol{\beta}|\boldsymbol{b}, \boldsymbol{z})$, $f(\boldsymbol{b}_i|\boldsymbol{\beta}, \boldsymbol{z}, \sigma_i^2)$, $f(\sigma_i^2|\boldsymbol{b}_i)$, $f(\boldsymbol{y}|\boldsymbol{z}, \gamma_2)$, and $f(\boldsymbol{z}, \gamma_2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b})$. We simulate from the joint distribution of $(\boldsymbol{z}, \gamma_2)$ to improve convergence of the overall joint distribution as suggested by Cowles [24]. Full details regarding our MCMC algorithm are described in the Appendix.

3.3. Calibration using multiple imputation

Imputation is used to correct for rater bias by replacing censored diagnoses (i.e. positive diagnoses by low-threshold raters and negative diagnoses by high-threshold raters) with imputed values. To preserve uncertainty due to censoring and incorporate parameter uncertainty, multiple imputations [25] are generated for each censored value.

Once the MCMC algorithm has achieved convergence, multiple imputations are created by drawing values of $\boldsymbol{z}$ and γ_2 from every kth iteration of one Markov chain (after a suitable burn-in period) and then applying the rules described in equation (3) to those participants with censored diagnoses. The size of the lag between iterations of the Markov Chain is chosen so that imputations are essentially independent. Once the censored values have been imputed, three calibration strategies can be employed:

1. Preserve the moderate category.
2. Collapse the moderate category into the severe category to produce overall diagnosis rates similar to those of the low-threshold raters.
3. Collapse the moderate category into the non-case category to produce overall diagnosis rates similar to those of the high-threshold raters.

Choice of the number of imputations is based on λ, the fraction of missing diagnosis information. Following [26], the asymptotic efficiency of an estimate based on m imputations relative to an infinite number of imputations is $(1+\lambda/m)^{-1/2}$ in units of standard deviations. If we assume that censored diagnoses are missing, then a conservative (since diagnoses are censored rather than missing and we also use covariates to predict diagnosis) estimate of the fraction of missing diagnosis information is the proportion of diagnoses that are censored. The number of imputations should be chosen so that the asymptotic efficiency of a post-imputation estimate is close to 1. Graham *et al.* [27] note that the rate in power falloff due to insufficient imputations is greater than that of relative efficiency and recommend between 20 and 40 imputations in most situations.

4. Application

We fit the model described in Section 3 to the WECare data using the MCMC algorithm described in the Appendix, which was programmed in R (code available from the first author). A weakly informative prior [28] was used for the random effect variances for both nurses and psychologists. Following Gelman [28], we chose priors that produced a posterior distribution consistent with external knowledge; in particular, we sought to preserve between-rater type variability and the PTSD diagnosis rates across nurses and psychologists (see Section 4.2). We used a Uniform (0,3) prior for both the nurse random effect standard deviation and the psychologist random effect standard deviation, which allows the standard deviation to span the full range of the latent data. To check that the choice of these priors was not distorting our inferences, we plotted a histogram of the posterior random effect standard deviations and overlaid the prior distribution.

We assessed convergence of our Markov chains by visual inspection of trace plots and autocorrelation plots of parameters as well as by using two formal convergence tests: univariate and multivariate potential-scale reduction factors [29] based on 10 Markov chains, and the diagnostic suggested by Geweke [30]. Autocorrelations plots and convergence tests were generated using the R package CODA [31].

4.1. Results from imputation model

Table III lists the variables and presents the posterior means and 95 per cent highest posterior density intervals for the regression coefficients in the WECare imputation model. We used baseline covariates (none of which had missing values) to generate the imputations. Following the advice of Rubin [32], we included all possibly relevant predictors of PTSD. The only significant variable in the imputation model is the number of stressful life events. This is partly due to the fact that many of the covariates in the imputation model are correlated with one another, reducing the significance of the individual regression coefficients.

4.2. Checking the fit of the model

Gelman *et al.* [33] recommend the use of posterior predictive checking (PPC) to investigate the fit of Bayesian models. With PPC, observed data are compared with the posterior predictive distribution of replicated data under the posited

Table III. Posterior means and 95 per cent highest posterior density intervals of regression coefficients from the WECare imputation model.

Regression coefficient	Mean	Lower 95 per cent HPD interval	Upper 95 per cent HPD interval
Intercept	0.012	−0.841	0.818
Baseline depression	0.013	−0.027	0.049
Baseline anxiety	0.004	−0.025	0.033
Raped	0.267	−0.095	0.590
Number of stressful life events	0.102	0.021	0.182
Age	−0.011	−0.033	0.010
Number of children	0.000	−0.125	0.117
Has a partner	0.060	−0.252	0.390
High school or more	0.040	−0.324	0.382
White	0.104	−0.523	0.772
Latina	0.080	−0.352	0.506

HPD: highest posterior density.

model. If the posited model accurately represents the process that generated the data, then replicated data generated under the model should look similar to the observed data. PPC is usually implemented by drawing predicted values from simulated values of the posterior distribution. The PPC is the comparison of the observed data to these predicted values using test statistics that characterize important features of the data. Discrepancies between the test statistics for the observed and replicated data indicate model misfit.

More specifically, following Gelman *et al.* [33], let y be the observed data, $p(\theta|y)$ be the joint posterior distribution of the parameters in the model, and define y^{rep} as the replicated data that could have been observed under our model that produced y. Then the posterior predictive distribution of y^{rep} is

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta, y)p(\theta|y)\,\mathrm{d}\theta \tag{5}$$

Test statistics for the observed data $T(y, \theta)$ are compared with those from the replicated data $T(y^{\text{rep}}, \theta)$ either using graphical plots or by the tail area probability p_B

$$p_B = \int\int I_{T(y^{\text{rep}},\theta)\leqslant T(y,\theta)} p(y^{\text{rep}}|\theta)p(\theta|y)\,\mathrm{d}y^{\text{rep}}\,\mathrm{d}\theta \tag{6}$$

where I is the indicator function. The quantity p_B is referred to as the posterior predictive p-value [33, p. 162].

We computed the posterior predictive distribution by simulation. Using the 10 000 simulations from our posterior density θ from one Markov chain, we discarded the first 2000 iterations and drew one y^{rep} from the predictive distribution for each simulated θ. This resulted in 8000 predictive test quantities $T(y^{\text{rep}}, \theta)$ that we compared with the realized test quantity $T(y, \theta)$. The posterior predictive p-value was calculated as the proportion of the 8000 simulations for which $T(y^{\text{rep}}, \theta)$ equals or exceeds $T(y, \theta)$. Two-sided p-values are generated by multiplying this proportion by 2 [34].

We choose four test statistics, which we felt summarized important features of the data that our model should capture. For both nurses and psychologists, we calculated the PTSD diagnosis rate and the between-rater standard deviation. Between-rater standard deviation for rater type i is calculated as $\sqrt{1/(m_i-1)\sum_{j=1}^{m_i} n_{ij}(\bar{y}_{ij}-\bar{y}_i)^2}$, where m_i is the number of raters of type i and n_{ij} is the number of patients rated by rater j of type i.

We choose between-rater standard deviation as one of our test quantities in order to verify that our model was accurately incorporating the fact that the between-rater variability of the psychologists was larger than the between-rater variability of the nurses. For this test quantity, we converted PTSD to a binary variable by collapsing the moderate and severe categories in equation (3) for nurses and the absence of condition and moderate condition categories for psychologists.

Graphical posterior checks of our four test statistics are presented in Figure 2. For all four test statistics, our model generates predicted results consistent with the observed data in the study; that is, the actual observations are typical of the predicted observations generated by the model. Two-sided posterior predictive p-values are all greater than 0.17.

4.3. Calibrating PTSD diagnoses using multiple imputation

To correct for the rater bias in the PTSD diagnoses, we replaced the censored PTSD diagnoses (i.e. positive diagnoses by nurses and negative diagnoses by psychologists) with imputed values using the rules described in equation (3). To

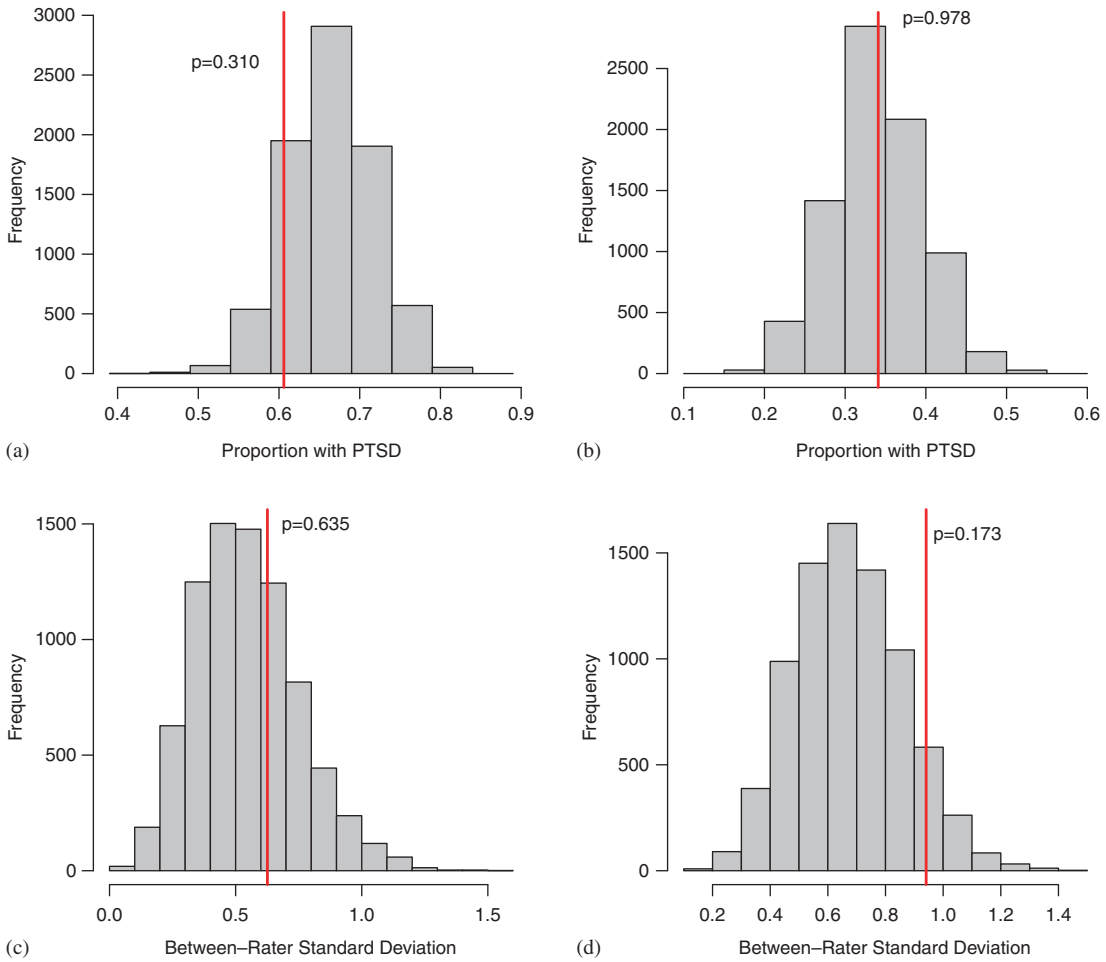

Figure 2. Observed data (indicated by a vertical line) vs data from 8000 simulated values from the posterior predictive distributions for four test quantities: (a) PTSD diagnosis rate of nurses (observed value at 0.606, two-sided p-value=0.310); (b) PTSD diagnosis rate of psychologists (observed value at 0.341, two-sided p-value=0.978); (c) between-rater standard deviation of nurses (observed value at 0.626, two-sided p-value=0.635); and (d) between-rater standard deviation of psychologists (observed value at 0.941, two-sided p-value=0.173). For all four test statistics, our model generates predicted results consistent with the observed data in the study; that is, the actual observations are typical of the predicted observations generated by the model.

preserve uncertainty due to censoring and incorporate parameter uncertainty, we created 20 multiple imputations [25] for each censored value.

If we assume that positive diagnoses by nurses and negative diagnoses by psychologists are missing, then an estimate of the fraction of missing information for PTSD is 48 per cent. Thus, for our data, an estimate based on $m=20$ imputations would tend to have a standard error only $(1+0.48/20)^{1/2}=1.01$ times as large as an estimate with an infinite number of imputations [26]. Graham *et al.* [27] show that in settings where the fraction of missing information is less than 50 per cent, 20 imputations provide minimal power and efficiency falloff as compared with 100 imputations. Therefore, 20 imputations is adequate in this setting.

Imputations were generated using the simulated parameters from one Markov chain of 10 000 iterations, discarding the first 2000 iterations and then using values of z and γ from every 400th iteration to generate 20 imputations for each censored PTSD diagnosis. Examination of autocorrelation plots determined that a lag of 400 iterations was large enough to ensure that our multiple imputations were generated from realizations of parameter values that could be regarded as independent [35].

Table IV tabulates PTSD status by rater type using imputed values. In addition to the original diagnoses, we present the data using three imputation-based approaches:

1. Preserve the moderate PTSD category.
2. Collapse the moderate PTSD category into the severe category to produce overall PTSD diagnosis rates similar to those of nurses.

Table IV. Imputed PTSD diagnosis percentages by rater type using three calibration approaches: (1) preserve the moderate PTSD category; (2) combine the moderate and severe PTSD categories; and (3) combine the moderate and non-PTSD categories.

Calibration approach	PTSD category	Nurse rate (per cent)	Psych rate (per cent)	Overall rate (per cent)	Test of independence
None. Use original diagnoses	Non-PTSD	39.4	65.9	52.3	$\chi^2 = 18.67$
	PTSD	60.6	34.1	47.7	$p < 0.0001$
Preserve moderate category	Non-PTSD	39.4	27.9	33.8	
	Moderate PTSD	32.4	38.0	35.1	$F_{2,179} = 1.54$
	Severe PTSD	28.2	34.1	31.1	$p = 0.7828$
Combine moderate and non PTSD category	Non-PTSD	71.8	65.9	69.0	$F_{1,450} = 0.93$
	PTSD	28.2	34.1	31.1	$p = 0.6647$
Combine moderate and severe category	Non-PTSD	39.4	27.9	33.8	$F_{1,117} = 2.42$
	PTSD	60.6	72.1	66.2	$p = 0.8775$

All three approaches resulted in a chi-squared test of independence for imputed data that was not significant, suggesting that rater bias had been alleviated. Twenty multiple imputations were generated for each censored PTSD diagnosis.

3. Collapse the moderate PTSD category into the non-PTSD category to produce overall PTSD diagnosis rates similar to those of psychologists.

The percentages in Table IV are column percentages identifying the PTSD diagnosis rates of nurses and psychologists across all 20 imputed values. In order to determine whether the methods have achieved balanced rates of diagnosis across rater types, we conducted tests of independence. The F statistic and p-value are from an approximation of the likelihood ratio test for independence using imputed data [36].

Looking at Table IV, all three calibration approaches produce results that do not reject the test for independence. When the moderate category is preserved, approximately one-third of participants fall into each PTSD category. When the moderate category is combined with the non-PTSD category, the nurse PTSD rate decreases from 60.6 to 28.2 per cent, which is very close to the psychologist PTSD rate of 34.1 per cent using the original PTSD diagnosis. This finding is consistent with our assumption that nurses tended to be lenient when diagnosing PTSD so that participants with moderate symptoms were classified as having PTSD.

Similarly, when the moderate category is combined with the severe category, the psychologist PTSD rate increases from 34.1 to 72.1 per cent, which is close the nurse PTSD rate of 60.6 per cent using the original PTSD diagnosis. This finding is consistent with our assumption that psychologists tended to be stringent when diagnosing PTSD so that participants with moderate symptoms were classified as not having PTSD.

4.4. Reanalysis of the WECare data using imputed PTSD scores

In this section, we conduct an analysis of the effects of comorbid PTSD on change in depression score in the WECare study. We focus on the first six months of follow-up since most change took place during this interval. In this analysis, PTSD is a covariate in the model and we are interested in the moderating effect of PTSD on depression treatment.

In addition to using the original observed diagnoses, we repeat the analysis using our imputed diagnoses based on the three different calibration approaches described in Table IV. Each calibration approach makes a different assumption regarding the censoring. Preserving the moderate category treats PTSD as an ordinal variable. Collapsing the moderate and severe categories assumes that the rate of PTSD determined by the nurses is the correct prevalence. Collapsing the moderate and non-PTSD categories assumes that the rate of PTSD determined by the psychologists is the correct prevalence. We analyze the data using a mixed-effects linear regression model. For subject i at interview j, $j = 1, \ldots, 6$, this model is

$$\begin{aligned}\text{Hamilton}_{ij} = {} & \beta_0 + \gamma_{i0} + \beta_1 \text{month}_{ij} + \gamma_{i1} \text{month}_{ij} + \beta_2 \text{month}_{ij}^2 + \beta_3 \text{MEDS}_i + \beta_4 \text{CBT}_i + \beta_5 \text{PTSD}_i + \beta_6 \text{MEDS}_i * \text{month}_{ij} \\ & + \beta_7 \text{CBT}_i * \text{month}_{ij} + \beta_8 \text{PTSD}_i * \text{month}_{ij} + \beta_9 \text{MEDS}_i * \text{month}_{ij} * \text{PTSD}_i + \beta_{10} \text{CBT}_i * \text{month}_{ij} * \text{PTSD}_i \\ & + \beta_{11} \text{WHITE}_i + \beta_{12} \text{LATINA}_i + \beta_{13} \text{Hamilton0}_i + \beta_{14} \text{Hamilton0}_i * \text{month}_{ij} + \varepsilon_{ij} \end{aligned} \quad (7)$$

where γ_{i0} and γ_{i1} are jointly normal correlated random intercept and slope parameters with mean 0, and month_{ij} indicates number of months since baseline for subject i at interview j. MEDS_i is a binary variable which indicates if subject i was assigned to medication, and CBT_i indicates whether subject i was assigned to CBT. WHITE_i and LATINA_i are

Table V. Estimated coefficients of treatment by month by PTSD interactions (and standard errors) from four mixed-effects regression models measuring change in depression over time.

Regression variable	(1) Original diagnosis	(2) Combine moderate and severe PTSD	(3) Combine moderate and non-PTSD	(4) Preserve moderate PTSD
MEDS by month by PTSD	0.178 (0.444)	−0.029 (0.502)	0.319 (0.637)	0.236 (0.653)
MEDS by month by moderate PTSD				−0.227 (0.609)
CBT by month by PTSD	0.484 (0.455)	0.385 (0.591)	0.427 (0.564)	0.554 (0.670)
CBT by month by moderate PTSD				0.246 (0.681)

Each model is based on a different PTSD calibration: (1) the original (biased) PTSD diagnoses; (2) imputed PTSD diagnoses where moderate and severe PTSD have been combined; (3) imputed PTSD diagnoses where moderate and non-PTSD have been combined; and (4) imputed PTSD diagnosis that preserves the moderate PTSD category.

binary variables indicating whether subject i is White or Latina, respectively; the reference group is African-Americans. Hamilton0_i is the baseline depression score for subject i, and the ε_{ij}'s are independent normal random error terms that are independent of the random intercept and slope. The variable PTSD$_i$ is an indicator variable that identifies whether the subject had PTSD at baseline. For the model that uses PTSD with three levels, PTSD$_i$ is replaced with two indicator variables, one variable indicating severe PTSD and the other indicating moderate PTSD. This model was fit separately on each imputed data set using the SAS procedure PROC MIXED [37] and estimates across the 20 multiple imputations were combined using the multiple imputation combining rules described in Rubin [25], implemented in the SAS procedure PROC MIANALYZE [37].

Inference focuses on the three-way treatment by month by PTSD interactions represented by the coefficients β_9 and β_{10} in equation (7). The WECare investigators hypothesized that CBT participants with comorbid PTSD would improve less over time compared with CBT participants without comorbid PTSD. They also hypothesized that there would be no difference in the slopes of medication participants with PTSD and those medication participants without PTSD since paroxetine is FDA-approved for PTSD.

Table V displays the results using the original diagnoses and three different calibration approaches. Although none of the parameter estimates are significant, their sign and magnitude are informative. Looking first at column 1 for original diagnoses, there is a positive medication by month by PTSD interaction effect and a large positive CBT by month by PTSD interaction effect. The large CBT by month by PTSD interaction is consistent with the WECare hypothesis that CBT would be less effective for participants with PTSD since the CBT intervention was not designed to treat PTSD. The positive medication by month by PTSD effect is contrary to the WECare hypothesis that this effect would be close to zero since paroxetine is FDA-approved to treat both depression and PTSD.

The results using imputed scores are more consistent with the WECare hypotheses. Looking at column 2 that combines moderate and severe PTSD into one category, we see a medication by month by PTSD effect which is close to zero suggesting that there was no difference in the effect of medication for participants with and without PTSD. The CBT by month by PTSD effect is still large and positive and consistent with the hypothesis that CBT is less effective for participants with PTSD.

Column 3 also uses imputed scores and combines moderate and non-PTSD into one category. Here, the CBT three-way interaction is similar to the previous columns but the medication three-way interaction is large again. Column 4, where imputed PTSD is kept on an ordinal scale, adds insight into what is driving the interactions in columns 2 and 3. For medication participants with severe PTSD, the interaction in column 4 is positive, suggesting that medication is less effective among those with severe PTSD. However, the medication by month by moderate PTSD interaction is negative, suggesting that medication is effective for those participants with moderate PTSD. For CBT participants, CBT becomes less effective as PTSD becomes more severe. Comparing the estimates in column 4 to those in column 2, we see that when we collapse the moderate and severe categories, the resulting parameter estimates fall between the column 4 estimates for moderate PTSD and those severe PTSD. When a PTSD diagnosis only consists of those with severe PTSD, as in column 3, the medication interaction is positive because medication is least effective among those with severe PTSD.

5. Discussion

We have proposed a framework for calibrating diagnoses that are subject to rater leniency bias and using the calibrated diagnoses in subsequent analyses. Our approach, developed in the context of two types of raters with different diagnosis thresholds, assumes an unobserved moderate category of patient that is assigned a positive diagnosis by one type of rater and a negative diagnosis by the other type. A Bayesian random effects censored ordinal probit model is used to calibrate the diagnoses across rater types by identifying and multiply imputing 'case' or 'non-case' status for patients in the moderate category, or preserving this category. Using multiple imputation for the diagnosis variable allows this variable to be used in a wide variety of subsequent analyses while also preserving uncertainty in true diagnosis. See the papers by Demirtas and Schafer [38, 39] and Demirtas [39] for other examples of Bayesian multiple imputation.

For the post-imputation analysis described in Section 4.4, a PTSD imputation model would ideally also condition on depression outcomes to preserve any relationship between PTSD and subsequent depression. However, we felt that most of the association between PTSD and depression was captured by conditioning on baseline depression in our imputation model. In addition, since depression outcomes had missing values, an imputation model that also included depression outcomes would necessitate the development of an imputation model that multiply imputed the diagnoses as part of a larger multiple imputation model that also imputes other variables with missing values.

We illustrated three alternative approaches for calibrating diagnoses: (1) combine the moderate and non-case categories; (2) combine the moderate and severe categories; (3) preserve the moderate category. The choice of which approach to use will depend on assumptions made by the analyst. In the WECare example, if one felt that the threshold used by the nurses was correct, then calibrating the diagnoses by combining the moderate and severe diagnoses would be the right approach. Conversely, if one felt that the threshold used by the psychologists was correct, then one would calibrate by combining the moderate and non-PTSD categories. We imagine that in many situations it will be clear to the investigator which group is using the correct threshold. In those situations where it is not clear which threshold is correct, preserving the moderate category and treating the diagnosis as an ordinal variable is an option—especially if the analyst is interested in the effects of a moderate category. Additional insights can be obtained using all three calibration methods and with little effort since they are all based on the same set of imputations. In the WECare example, all three calibration approaches gave results that differed from the analysis using the original unbalanced diagnoses.

Many extensions of our methods are possible. Our application involved two rater types, requiring two latent thresholds. The setting of more than two rater types could be accommodated by modeling additional thresholds. We used diffuse priors and assumed that the random effects in our model followed a normal distribution. Alternative priors and a different random effect distribution may do a better job calibrating diagnosis rates in other settings. We used the calibrated diagnosis variable as a covariate in regression analyses. If diagnosis is to be used as an outcome variable, then one could fit a mixed-effects logistic or probit regression model with heterogeneous random rater-specific intercepts and fixed rater-type effects [40], which would control the biased ratings. While we used a frequentist approach for our end-analysis, the proposed method can also be used as part of a fully Bayesian approach. In this context, multiple imputation can be viewed as a computational device for approximating a fully parametric Bayesian analysis (see, e.g. [41]). One could also use the simulated draws from the posterior distribution of the model in Section 3.2 to estimate parameters of interest. For example, if one were interested in a calibrated prevalence of PTSD, then one could simply average the simulated PTSD value from the MCMC iterations.

Our methods have potential application to several areas of medicine. In general, any assessments that are inferential and involve raters who differ on important characteristics such as experience or educational background are vulnerable to rater bias. Some possible areas for application of our methods include mammogram screening by radiologists [42], attention deficit hyperactivity disorder assessments by parents and teachers [43], diagnoses of dementia [44], other applications of PTSD [16, 45], and comparisons between two different types of diagnostic interviews [46].

Appendix A: MCMC algorithm

Here we describe the algorithm for drawing from the joint posterior distribution $[\theta|\text{data}]$ using MCMC simulation methods. The algorithm is a hybrid Metropolis–Hastings/Gibbs Sampler that follows closely to those of Xie *et al.* [22] and Johnson and Albert [21, chapter 4] and cycles between the following conditional distributions: $f(\boldsymbol{\beta}|\boldsymbol{b},\boldsymbol{z})$, $f(\boldsymbol{y}|\boldsymbol{z},\gamma_2)$, $f(\boldsymbol{z},\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})$, $f(b|\boldsymbol{\beta},\boldsymbol{z},\sigma_i^2)$, and $f(\sigma_i|\boldsymbol{b})$. We discuss each of these conditional distributions in turn.

A.1. Simulating from $f(\boldsymbol{\beta}|\boldsymbol{b},\boldsymbol{z})$

Assuming a uniform prior, the conditional distribution of the regression vector $\boldsymbol{\beta}$ given $\boldsymbol{b}$, $\boldsymbol{z}$, and covariate matrix $\mathbf{X}$ is

$$f(\boldsymbol{\beta}|\boldsymbol{b},\boldsymbol{z}) \sim \mathrm{N}((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{z}-\boldsymbol{b}),(\mathbf{X}'\mathbf{X})^{-1})$$

A.2. Simulating from $f(y_{ijk}|z_{ijk},\gamma_2)$

Values of y_{ijk} are drawn using equation (3).

A.3. Simulating from $f(\boldsymbol{z},\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})$

We factor this density into the product,

$$f(\boldsymbol{z},\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})=f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b},\gamma_2)f(\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})$$

and then simulate from $f(\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})$ using a Metropolis–Hastings step as described by Cowles [24] and simulate from $f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b},\gamma_2)$ using the truncated normal distributions described below.

To simulate from $f(\gamma_2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b})$ using a Metropolis–Hastings step, first generate the candidate threshold g_2 for updating γ_2 by drawing from $g_2\sim \mathrm{N}(\gamma_2,\sigma^2_{\mathrm{MH}})$ truncated to $(0,\infty)$. An appropriate value for σ_{MH} is chosen to obtain an acceptance rate of approximately 0.44, which Gelman *et al.* [47] found to be optimal for univariate Metropolis–Hastings chains of certain types. For the WECare data, we found that a value of $\sigma_{\mathrm{MH}}=0.14$ provides an acceptance rate of approximately 0.53. With this candidate generating density, the acceptance probability α is equal to $\min(1,r)$ where

$$r=\frac{\Phi(\gamma_2/\sigma_{\mathrm{MH}})}{\Phi(g_2/\sigma_{\mathrm{MH}})}\prod_{y_{ijk}\in 2}\frac{\Phi(\gamma_2-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})-\Phi(-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})}{\Phi(g_2-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})-\Phi(-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})}\times\prod_{y_{ijk}\in 3}\frac{1-\Phi(\gamma_2-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})}{1-\Phi(g_2-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}-b_{j(i)})}$$

The conditional distribution for $f(z_{ijk}|\gamma_2,y_{ijk},\boldsymbol{\beta},b_{j(i)})$ is

$$z_{ijk}|\gamma_2,y_{ijk},\boldsymbol{\beta},b_{j(i)}\sim\begin{cases}\phi(\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}+b_{j(i)},1) & \text{truncated to the interval } (-\infty,0) \quad \text{if } y_{ijk}=0\\ \phi(\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}+b_{j(i)},1) & \text{truncated to the interval } (\gamma_2,\infty) \quad \text{if } y_{ijk}=2\\ \phi(\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}+b_{j(i)},1) & \text{truncated to the interval } (-\infty,\gamma_2) \quad \text{if } k\in\mathscr{L}\\ \phi(\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta}+b_{j(i)},1) & \text{truncated to the interval } (0,\infty) \quad \text{if } k\in\mathscr{R}\end{cases}$$

where as before, $\mathscr{R}$ is the set of participants who received a positive diagnosis from a low-threshold rater and $\mathscr{L}$ is the set of participants who received a negative diagnosis from a high-threshold rater.

A.4. Simulating from $f(\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{z},\sigma_i^2)$

Assuming a normal prior on $\boldsymbol{b}_i$, the conditional distribution of $b_{j(i)}$ given $\boldsymbol{\beta}$, $\boldsymbol{z}$, and σ_i^2 is

$$b_{j(i)}|\boldsymbol{\beta},z_{ijk},\sigma_i^2\sim\mathrm{N}(\hat{b}_{j(i)},n_{ij}+\sigma_i^{-2})$$

where $\hat{b}_{j(i)}=(n_{ij}+\sigma_i^{-2})\times(\sum_{k=1}^{n_{ij}}(z_{ijk}-\boldsymbol{x}_{ijk}{}^{\mathrm{T}}\boldsymbol{\beta})$, and n_{ij} is the number of participants rated by rater j of rater type i.

A.5. Simulating from $f(\sigma_i|\boldsymbol{b}_i)$

Following Gelman [28], we assume that σ_i has Uniform $(0,q)$ prior. Then, the conditional distribution $f(\sigma_i|\boldsymbol{b}_i)$ is proportional to

$$\sigma_i|b_{j(i)}\propto\sigma_i^{-m_i}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}b_{j(i)}^2\right),\quad 0<\sigma_i<q \tag{A1}$$

where m_i is the number of raters of type i. We sample from this distribution by setting up a grid for σ_i, evaluating equation (A1) at each grid point, and then drawing from this discrete approximation to the posterior distribution of σ_i. The grid here is 2000 points equally spread from 0 to q.

Acknowledgements

The authors wish to thank Jeanne Miranda for the WECare data and C. Hendricks Brown, the Prevention Sciences Methodology Group (R01MH040859), and three anonymous reviewers whose comments greatly improved the quality of this manuscript. This work was partially funded by a seed grant from the Center for Health Administration Studies while Dr Siddique was a Research Associate at the University of Chicago. Dr Crespi was funded by NIH grant CA16042. This work is supported by NIH grant R01MH066302.

References

1. Hoyt WT, Kerns MD. Magnitude moderators of bias in observer ratings: a meta-analysis. *Psychological Methods* 1999 **4**:403–424. DOI: 10.1037/1082-989X.4.4.403.
2. Kneeland N. That lenient tendency in rating. *Personnel Journal* 1929; **7**:356–366.
3. Hoyt WT. Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods* 2000; **5**:64–86. DOI: 10.1037/1082-989X.5.1.64.
4. Raymond MR, Viswesvaran C. Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement* 1993; **30**:253–268. DOI: 10.1111/j.1745-3984.1993.tb00426.x.
5. de Gruijter DN. Two simple models for rater effects. *Applied Psychological Measurement* 1984; **8**:213–218. DOI: 10.1177/014662168400800210.
6. Houston WM, Raymond MR, Svec JC. Adjustments for rater effects in performance assessment. *Applied Psychological Measurement* 1991; **15**:409–421. DOI: 10.1177/014662169101500411.
7. Braun HI. Understanding scoring reliability: experiments in calibrating essay readers. *Journal of Educational and Behavioral Statistics* 1988; **13**(1):1–18. DOI: 10.3102/10769986013001001.
8. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Health Disorders* (4th edn). American Psychiatric Association: Washington, DC, 1994.
9. Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, severity, comorbidity of twelve-month DSM-IV disorders in the national comorbidity survey replication (NCS-R). *Archives of General Psychiatry* 2005; **62**:617–627. DOI: 10.1001/archpsyc.62.6.617.
10. Davidson JR. Trauma: the impact of post-traumatic stress disorder. *Journal of Psychopharmacology* 2000; **14**:S5–S12.
11. Margolin G, Gordis EB. The effects of family and community violence on children. *Annual Review of Psychology* 2000; **51**:445–479. DOI: 10.1146/annurev.psych.51.1.445.
12. Yehuda R. Biological factors associated with susceptibility to posttraumatic stress disorder. *Canadian Journal of Psychiatry* 1999; **44**:34–39.
13. Regier DA, Rae DS, Narrow WE, Kaelber CT, Schatzberg AF. Prevalence of anxiety disorders and their comorbidity with mood and addictive disorders. *British Journal of Psychiatry Supplement* 1998; **34**:24–28.
14. Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS, Keane TM. The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress* 1995; **8**:75–90. DOI: 10.1002/jts.2490080106.
15. Norris FH, Hamblen JL. Standardized self-report measures of civilian trauma and PTSD. In *Assessing Psychological Trauma and PTSD*, Wilson JP, Keane TM, Martin T (eds). Guilford Press: New York, 2004; 63–102.
16. Breslau N, Kessler R, Peterson EL. Post-traumatic stress disorder assessment with a structured interview: reliability and concordance with a standardized clinical interview. *International Journal of Methods in Psychiatric Research* 1998; **7**:121–127. DOI: 10.1002/mpr.41.
17. Miranda J, Chung JY, Green BL, Krupnick J, Siddique J, Revicki DA, Belin T. Treating depression in predominantly low-income young minority women. *Journal of the American Medical Association* 2003; **290**:57–65. DOI: 10.1001/jama.290.1.57.
18. Spitzer RL, Williams JB, Gibbon M, First MB. *Structured Clincal Interview for DSM-III-R–Patient Edition* (*with Psychotic Screen*)-*SCID-P*, American Psychiatric Press: Washington, DC, 1990.
19. Williams JB. A structured interview guide for the Hamilton depression rating scale. *Archives of General Psychiatry* 1988; **45**:742–747. DOI: 10.1001/archpsyc.1988.01800320058007.
20. Green BL, Krupnick JL, Chung J, Siddique J, Krause ED, Revicki D, Frank L, Miranda J. Impact of PTSD comorbidity on one-year outcomes in a depression trial. *Journal of Clinical Psychology* 2006; **62**:815–835. DOI: 10.1002/jclp.20279.
21. Johnson VE, Albert JH. *Ordinal Data Modeling*. Springer: New York, 1999.
22. Xie M, Simpson DG, Carroll RJ. Random effects in censored ordinal regression: latent structure and Bayesian approach. *Biometrics* 2000; **56**:376–383. DOI: 10.1111/j.0006-341X.2000.00376.x.
23. Gelfand AE, Smith AF. Gibbs sampling for marginal posterior expectations. *Communications in Statistics*, *Theory and Methods* 1991; **20**:1747–1766. DOI: 10.1080/03610929108830595.
24. Cowles MK. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing* 1996; **6**:101–111. DOI: 10.1007/BF00162520.
25. Little RJ, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
26. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
27. Graham JW, Olchowski AE, Gilreath TD. How man imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 2007; **8**:206–213. DOI: 10.1007/s11121-007-0070-9.
28. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**:515–534. DOI: 10.1214/06-BA117A.
29. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**:434–455. DOI: 10.2307/1390675.
30. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, Bernardo JM, Berger JO, Dawid AP, Smith AF (eds). Oxford University Press: Oxford, 1992;169–193.
31. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News* 2006; **6**:7–11.
32. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489. DOI: 10.2307/2291635.
33. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall/CRC: New York, 2004.
34. Gelman A, Goegebeur Y, Tuerlinckx F, Mechelen IV. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*) 2000; **49**:247–268. DOI: 10.1111/1467-9876.00190.
35. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC: New York, 1997.
36. Li KH, Raghunathan TE, Rubin DB. Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* 1991; **86**:1065–1073.
37. SAS Institute Inc. *The SAS System*, Version 9.1. SAS Institute Inc., Cary, NC, 2003.
38. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2003; **22**:2553–2575. DOI: 10.1002/sim.1475.
39. Demirtas H. Multiple imputation under bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2005; **24**:2345–2363. DOI: 10.1002/sim.2117.

40. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Wiley: New York, 2006.
41. Dominici F, Zeger S, Parmigiani G, Katz J, Christian P. Estimating percentile-specific treatment effects in counterfactual models: a case study of micronutrient supplementation, birth weight and inflant mortality. *Journal of the Royal Statistical Society Series C Applied Statistics* 2006; **55**:261–280. DOI: 10.1111/j.1467-9876.2006.00533.x.
42. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, Fletcher SW. Screening mammograms by community radiologists: variability in false-positive rates. *Journal of the National Cancer Institute* 2002; **94**:1373–1380.
43. Swanson JM, Hinshaw SP, Arnold LE, Gibbons RD, Marcus S, Hur K, Jensen PS, Vitiello B, Abikoff HB, Greenhill LL, Hechtman L, Pelham WE, Wells KC, Conners CK, March JS, Elliott GR, Epstein JN, Hoagwood K, Hoza B, Molina BS, Newcorn JH, Severe JB, Wigal T. Secondary evaluations of MTA 36-month outcomes: propensity score and growth mixture model analyses. *Journal of the American Academy of Child and Adolescent Psychiatry* 2007; **46**:1003–1014.
44. Tractenberg RE, Schafer K, Morris JC. Interobserver disagreements on clinical dementia rating assessment: interpretation and implications for training. *Alzheimer Disease and Associated Disorders* 2000; **15**:155–161.
45. Kulka RA, Schlenger WE, Fairbank JA, Hough RL, Jordan BK, Marmar CR, Weiss DS. *The National Vietnam Veterans Readjustment Study*: *Tables of Findings and Technical Appendices*. Brunner/Mazel: New York, 1990.
46. Taub NA, Morgan Z, Brugha TS, Lambert PC, Bebbington PE, Jenkins R, Kessler RC, Zaslavsky AM, Hotz T. Recalibration methods to enhance information on prevalence rates from large mental health surveys. *International Journal of Methods in Psychiatric Research* 2005; **14**:3–13. DOI: 10.1002/mpr.13.
47. Gelman A, Roberts GO, Gilks WR. Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, Bernardo JM, Berger JO, Dawid AP, Smith AF (eds). Oxford University Press: Oxford, 1996; 599–607.