Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE

A Bayesian Approach to Modeling Variance of Intensive Longitudinal Biomarker Data as a Predictor of Health Outcomes

Mingyan Yu[1] | Zhenke Wu[1] | Margaret Hicken[2] | Michael R. Elliott[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA | [2]Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

Correspondence: Mingyan Yu (myanyu@umich.edu)

Received: 15 January 2024 | **Revised:** 26 July 2024 | **Accepted:** 27 October 2024

Funding: This work was supported by National Institutes of Health (R56AG-066693 and R01MD-016244).

Keywords: bayesian hierarchical models | intensive heart rate | intensive longitudinal biomarker | joint models | social stress | variability

ABSTRACT

Intensive longitudinal biomarker data are increasingly common in scientific studies that seek temporally granular understanding of the role of behavioral and physiological factors in relation to outcomes of interest. Intensive longitudinal biomarker data, such as those obtained from wearable devices, are often obtained at a high frequency typically resulting in several hundred to thousand observations per individual measured over minutes, hours, or days. Often in longitudinal studies, the primary focus is on relating the means of biomarker trajectories to an outcome, and the variances are treated as nuisance parameters, although they may also be informative for the outcomes. In this paper, we propose a Bayesian hierarchical model to jointly model a cross-sectional outcome and the intensive longitudinal biomarkers. To model the variability of biomarkers and deal with the high intensity of data, we develop subject-level cubic B-splines and allow the sharing of information across individuals for both the residual variability and the random effects variability. Then different levels of variability are extracted and incorporated into an outcome submodel for inferential and predictive purposes. We demonstrate the utility of the proposed model via an application involving bio-monitoring of hertz-level heart rate information from a study on social stress.

1 | Introduction

Over the past several decades a large literature has developed to use longitudinal measurements such as biomarkers or survey responses to predict health outcomes. For example, Henderson, Diggle, and Dobson [1] linked linear mixed effects models for longitudinal measures of psychiatric disorders with semi-parametric proportional hazard models for time-to-drop-out through a latent bivariate Gaussian process to explore the association between longitudinal psychiatric disorder scores and the outcome of interest. Wu, Hu, and Wu [2] extended the sub-model for longitudinal biomarkers into non-linear mixed effects models and connected them with the survival sub-model via subject-level random effects estimated from the longitudinal trajectories. Horrocks and van Den Heuvel [3] used longitudinal measurements of the adhesiveness of $CD56^{bright}$ cells to predict the probability of achieving a successful pregnancy.

The development of ecological momentary assessment (EMA) data [4], which is typically provided at much higher frequencies with consequently much larger individual-level sample sizes, has spawned a smaller but rapidly developing literature to

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

incorporate EMA data into the prediction of health outcomes. EMA data can occur in a variety of forms: for example, prompts on smartphones or other electronic devices to complete questionnaires or undertake tasks [5]; respondent tracking of events occurring in their experience [6]; or, as in our example, biometric measures such as skin conductivity or heart rates. By virtue of its density, EMA data allows both assessment of short term effects and the possibility of more accurate prediction or association with longer-term outcomes. But this density as well as the variety of ways in which this data is collected also may require special consideration of statistical methods for its analysis [7]. In general, some types of methods that account for the longitudinal nature of the data, such as ANOVA [5], or linear or generalized linear mixed models [6, 8], or structural equation modeling [9], have been used.

While a considerable proportion of studies relating longitudinal data to health outcomes have primarily focused on the mean structure of the longitudinal measurements and treated the variance as a nuisance parameter, it is crucial to realize that the variance structure itself can be of vital importance and can have a strong impact on the subsequent analysis [10]. The literature in this area is small but growing; for example, Palmier-Claus et al. [11] studied the association between affective variability and suicidal ideation by regressing suicidal ideation scores on daily variability of negative and positive affect scores calculated by mean squared successive difference of the self-reported affect scores within each day. Mood variability was found to have a significant and stronger association with the frequency and severity of suicidality than the daily average. Fang et al. [12] found that reduced average daily total sleep time along with increased variability in daily total sleep time were correlated with higher levels of depressive symptoms among first-year medical interns. They discovered this relationship by regressing the quarterly average depression scores on the mean and standard deviation of daily total sleep time collected over the first internship year through the Intern Health Study. Taylor et al. [13] concluded from the Study of Women's Health Across the Nation (SWAN) sleep study that greater variability in bedtime rather than later bedtime was associated with metabolic health in midlife women by relating the standard deviations of individuals' sleep time over 2 weeks to insulin resistance. Gao et al. [14] found that individual-level variability of intraocular pressure is an independent risk factor for onset of primary open-angle glaucoma. Elliott, Sammel, and Faul [15] linked mean and variance of cognitive scores to the development of dementia, finding that variability (but not mean) was predictive of dementia onset. Jiang et al. [16] considered a mixture model which included latent classes for the individual-specific mean profiles and residual variances, finding that higher variability in Follicle Stimulating Hormone (FSH) was associated with a higher chance of experiencing hot flashes in females.

Consideration of variability as a direct predictor in EMA data does not seem to exist. Hedeker, Mermelstein, and Demirtas [17] did model between- and within-subject variances of negative and positive affect obtained by random prompts from handheld computers as a function of baseline covariates via log-linear models, but did not consider variability of the affect measures themselves as predictors.

A key feature of EMA data is that it is obtained at a high frequency and typically results in hundreds to thousands of observations per individual taken over hours or days. Furthermore, here we focus on heart rate, which, due to autonomic regulation of the cardiovascular system, is approximately stationary except over very short time intervals, so that mean trends are not available as predictors. Heart rate variability (HRV) has been considered as a predictor of several health outcomes outside of EMA settings [18, 19]. The time and frequency domain of HRV are first derived based on the interbeat intervals, which are then averaged over a specific time period (e.g., 5-min average HRV). A previous study on HRV and stress have shown that standard deviations of interbeat intervals (SDNN), which measures the overall HRV, is significantly lower when people were under stress [20].

Instead of simple averaging, our approach to the analysis of heart rate data is different and, we believe, highly novel and likely to access quite a bit more information from standard HRV. First, we model variability in heart rates using a random effects B-spline, separating variance into a longer-term variability component measured by the spline random effects variance ("wiggliness" of the heart rate) and a shorter-term variability component measured by the residual variance ("second-to-second" variability of the heart rate). We also consider an individual autocorrelation measure [21] to account for correlation in residual variance. We use a joint model, with the individual level "wiggleness" and "second-to-second" variability modeling assignment to one of four controlled stress settings. This joint modeling is a key feature of our development as well since two-stage approaches that estimate variabilities of longitudinal measurements in the first stage and then plug in these sample-based estimates into the second stage outcome model will usually lead to biased inference and uncontrolled variance in the second stage model [22] as the uncertainty of the mean and variance estimation in the first stage is not properly propagated to the outcome model. Corrections have been proposed to eliminate the bias in two-stage models. For example, Ye, Lin, and Taylor [23] proposed a risk set regression calibration (RRC) method to correct the bias inherent in the naive two-stage model when connecting longitudinal measurements and the time-to-event outcome of interest; however, their method worked well only when measurement error in the longitudinal marker was low. Joint modeling approaches have attracted increasing attention in the past two decades and simulation studies have shown that joint modeling the longitudinal measurements and cross-sectional or survival outcomes together generally leads to unbiased results [24]. These joint models are often implemented using shared random effects models [25], where a first-stage model is used to estimate subject-level trends using random effects, and these random effects are then used as predictors in the second-stage models [26, 27]. We follow such an approach here.

We demonstrate the application of our work to the Michigan Work-Life Study where we aim to determine how variability in hertz-level heart rate measurements are associated with stress, which is reflected by the group assignments. This study was designed to examine the nature of race-based anticipatory and perseverative stress among Black women. Using the Trier Social Stress Test (TSST) to elicit a social stress response (analogous to the use of the treadmill to elicit a physical stress response), participants were randomly assigned to one of four race-related

stress exposure groups (described in detail below). The TSST includes three phases: (a) anticipation, which captures anticipatory stress; (b) social stress, which is the actual acute stress exposure; and (c) recovery, which may capture any perseverative stress. The biological stress responses during the TSST can be measured in numerous ways; in the Michigan Work-Life Study, these responses were measured using blood pressure, electrodermal activity, heart rate, salivary cortisol, and salivary alpha-amylase. In this applied demonstration of this novel statistical approach, we focus on hertz-level heart rate information. We take a non-standard approach in this analysis, in part for illustrative purposes of the method, treating the assignment group as the outcome. Thus our analysis will focus on understanding how variability in heart rate is associated with the various treatment assignments rather than viewing variability strictly as a predictor in a formal sense.

The remainder of our manuscript is organized as follows. Section 2 develops the statistical details of the proposed joint model in a Bayesian framework. Section 3 implements a simulation study to assess the quality of inference from the joint model, and compares this with somewhat simpler a two-stage approach. Section 4 applies our method to the Michigan Work-Life Study. Section 5 summarizes the findings of the simulation study and application and considers limitations and the next step for research.

2 | Methods

2.1 | Observed Data

Let Y_i denote the cross-sectional binary outcome of subject $i = 1, \ldots, N$. Let $\boldsymbol{X}_i = \left[X_{i1}, \ldots, X_{in_i}\right]^\top$ be a vector of n_i longitudinal biomarker values collected for subject i over time and $\boldsymbol{t}_i = \left[t_{i1}, \ldots, t_{in_i}\right]^\top$ be a vector of n_i corresponding time points. Let X_{ij} denote the value of the longitudinal biomarker of subject i collected at time t_{ij}, $j = 1, \ldots, n_i$. Baseline covariates are denoted by vector $\boldsymbol{Z}_i$.

2.2 | Joint Model

To jointly model the longitudinal biomarker and the binary outcome together, the proposed model is comprised of two components. In the longitudinal marker sub-model, we model the intensive longitudinal marker data $X_{ij}, i = 1, \ldots, N, j = 1, \ldots, n_i$ as a non-linear function of t_{ij} using a B-spline regression model. The outcome sub-model then links the short-term and long-term variances from the B-spline longitudinal marker sub-model with the binary outcome Y_i.

2.2.1 | Predictor Sub-Model: Longitudinal Biomarker

The proposed model for the longitudinal marker is as follows:

$$\begin{aligned} \boldsymbol{X}_{\boldsymbol{i}} &\sim \boldsymbol{N}_{n_i}\left(f\left(\boldsymbol{t}_{\boldsymbol{i}}; \beta_0, \boldsymbol{b}_{\boldsymbol{i}}\right), \boldsymbol{\Sigma}_{\boldsymbol{i}}\right), \quad \text{independently for } i = 1, \ldots, N \\ b_{il} &\overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2_{b_i}\right), \quad l = 1, \ldots, L \end{aligned} \tag{1}$$

Here, $\boldsymbol{N}_{n_i}\left(\boldsymbol{f}_i, \boldsymbol{\Sigma}_{\boldsymbol{i}}\right)$ denotes an n_i-dimensional multivariate normal distribution with mean $\boldsymbol{f}_i$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{i}}$, where n_i represents the number of observations for participant i. The function $f\left(\boldsymbol{t}_{\boldsymbol{i}}; \beta_0, \boldsymbol{b}_{\boldsymbol{i}}\right)$ represents a non-linear function of time $\boldsymbol{t}_i$ characterized by population-level fixed effects β_0 and individual-level random effects $\boldsymbol{b}_{\boldsymbol{i}}$. In our case, this non-linear function is a B-spline regression model provided by $f\left(t_{ij}; \beta_0, \boldsymbol{b}_{\boldsymbol{i}}\right) = \beta_0 + \sum_{l=1}^{L} b_{il}\xi_{l,d}\left(t_{ij}\right)$. The individual-level random effects $\boldsymbol{b}_{\boldsymbol{i}} = \left[b_{i1}, \ldots, b_{iL}\right]^\top$ is a vector of length L where L denotes the number of basis functions of time and we assume participants have the same number of basis functions here. $\xi_{l,d}(\cdot)$ is the B-spline basis function of degree d and is formulated by $\xi_{l,d}\left(t_{ij}\right) = \frac{t_{ij}-k_l}{k_{l+d}-k_l}\xi_{l,d-1}\left(t_{ij}\right) + \frac{k_{l+1+d}-t_{ij}}{k_{l+1+d}-k_{l+1}}\xi_{l+1,d-1}\left(t_{ij}\right)$ with $\xi_{l,0}\left(t_{ij}\right) = I\left(k_l \le t_{ij} \le k_{l+1}\right)$; k_l's in the formulation are the time knots based on specific selections. In our simulation and application, we set the degree of the B-spline basis function to be 3 and placed a time knot at every 2nd percentile of the observed time duration. Since intensive longitudinal measurements are likely to be correlated with each other, we can impose different covariance matrices to reflect the correlations. We can rewrite $\boldsymbol{\Sigma}_i$ as $\sigma^2_{e_i}\boldsymbol{R}_{\boldsymbol{i}}$, where we assume that observations for participant i have a constant residual variance $\sigma^2_{e_i}$ across time and $\boldsymbol{R}_{\boldsymbol{i}}$ is the correlation matrix.

This model uses individual-specific variability in the random effects, $\sigma^2_{b_i}$'s, to characterize "long-term" variability (i.e., minute-to-minute variability) in the longitudinal marker, and uses individual-specific variability in the residual errors, $\sigma^2_{e_i}$'s, to characterize "short-term" variability (i.e., second-to-second variability) in the longitudinal marker. The traditional definition of "Heart Rate Variability" is the variation in the duration between consecutive heart beats, also called an "interbeat interval," as the heart continuously adapts to the environment, and includes numerous time- and frequency-based components important in clinical settings [28]. Faster heart rates mean that the interbeat interval is shorter, with less opportunity for variation; slower heart rates mean that the interbeat interval is longer, allowing for more opportunity for variation in heart rate or greater HRV. Basic time-based measures of HRV are generally calculated as the standard deviation or the root mean square of successive interbeat intervals over a short epoch, yielding single average values of HRV. However, here, we examine a different approach to quantifying this time-based variability.

2.2.2 | Outcome Sub-Model: Binary Outcome

The binary outcome Y_i is assumed to be associated with the individual-specific variances and baseline covariates $\boldsymbol{Z}_i$ through the following probit regression model:

$$\begin{aligned} \boldsymbol{S}_i &= \left[\sigma_{b_i}, \sigma_{e_i}, \boldsymbol{Z}_i^\top\right]^\top \\ Y_i &\sim \text{Bern}\left(\Phi\left(\eta\left(\boldsymbol{S}_i\right)\right)\right) \\ \eta\left(\boldsymbol{S}_i\right) &= \alpha_1 + \alpha_2 \times \log\left(\sigma_{b_i}\right) + \alpha_3 \times \log\left(\sigma_{e_i}\right) + \boldsymbol{\gamma}\boldsymbol{Z}_i, i = 1, \ldots, N \end{aligned} \tag{2}$$

$\boldsymbol{S}_i$ is a vector of individual-specific short-term variability σ_{e_i}, long-term variability σ_{b_i} and baseline covariates $\boldsymbol{Z}_i$. The outcome is modeled using a logistic regression with a probit link; Φ in (2) denotes the cumulative density function (CDF) of

the standard normal distribution. $\eta(\boldsymbol{S}_i)$ is a linear function of log-transformed variances and baseline covariates. (In our example we do not include baseline covariates since the randomization of Y means that there was little imbalance with respect to covariates, but we include them in our model for completeness.) Although we specify a probit model for a binary outcome, a generalized linear model with a specified link function can be considered.

2.3 | Priors

We specify the following priors and hyperpriors for each unknown parameter.

For the longitudinal marker sub-model fixed effects β_0, we assume $\beta_0 \sim N(0, \psi^2)$.

To allow sharing of information across individuals for long-term variability, we propose a common prior distribution for $\sigma^2_{b_i}$'s

$$\log\left(\sigma^2_{b_i}\right) \sim N(\upsilon_b, \Psi^2_b) \tag{3}$$

with hyper-parameter priors [29]

$$\upsilon_b \sim N(m_b, \kappa^2_b), \Psi_b \sim \text{Half} - \text{Cauchy}(0, \tau_b) \tag{4}$$

A general form for the prior distribution for the subject-level correlation matrix $\boldsymbol{R}_i$ is difficult to specify for $n_i > 2$, given the need for $\boldsymbol{R}_i$ to be positive definite. Options to deal with this include positing independent $U(-1, 1)$ distribution for correlation parameters and rejecting those that fail to be jointly positive definite [30] and parameterization of correlations in terms of hyperspherical coordinates [31]. Here we elect to take advantage of a high degree of beat-to-beat autocorrelation and assume an AR1 model so that the off-diagonal elements of the correlation matrix $\boldsymbol{R}_i$ become $r_{i,ks} = \rho_i^{|s-k|}$ for $1 \le k, s \le n_i$ with ρ_i representing the autocorrelation parameter for individual i. We implement this model by decomposing the original error $\epsilon_{ij} = \rho_i \epsilon_{ij-1} + w_{ij}$ into a deterministic function of the error at the previous time point, $\rho_i \epsilon_{ij-1}$, and an innovative new "white noise" w_{ij}, where $w_{ij} \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2_{w_i}\right)$. We assume the priors for this as

$$\begin{aligned} \log\left(\sigma^2_{w_i}\right) &\sim N(\upsilon_w, \Psi^2_w) \\ \upsilon_w \sim N(m_w, \kappa^2_w), \Psi_w &\sim \text{Half} - \text{Cauchy}(0, \tau_w) \\ \rho_i &\sim U(0, 1) \end{aligned} \tag{5}$$

where the prior on ρ_i assumes a positive autocorrelation that is biologically warranted for heart rates. Recall that in Section 2.2.1, we decompose the covariance matrix of $\boldsymbol{X}_i$ into $\boldsymbol{\Sigma}_i = \sigma^2_{e_i} \boldsymbol{R}_i$, where we assume $Var(\epsilon_{ij}) = \sigma^2_{e_i}$. With the specification of ρ_i and the prior for w_{ij}, we naturally get the deterministic relationship $\sigma^2_{e_i} = \frac{\sigma^2_{w_i}}{1-\rho^2_i}$, which makes it unnecessary to specify a prior for $\sigma^2_{e_i}$.

For the outcome sub-model, we assume an independent $N(0, \omega^2)$ priors for each of the regression coefficients $(\alpha_1, \alpha_2, \alpha_3, \boldsymbol{\gamma})$.

For the hyperparameters, we chose values that correspond to weakly informative values. Thus for the fixed effect regression hyperparameters, we set $\psi = 1000$ and for the outcome sub-model regression hyperparameters, we set $\omega = 10$. For the variance hyperparamter model, we set $m_b = m_w = 0$, $\kappa_b = \kappa_w = 1000$, and $\tau_b = \tau_w = 2.5$.

2.4 | Joint Distribution

Let $\boldsymbol{\Theta} = \left[\beta_0, \upsilon_b, \Psi_b, \upsilon_w, \Psi_w, \rho_i, \alpha_1, \alpha_2, \alpha_3, \boldsymbol{\gamma}\right]^\top$ denote the vector of unknown parameters of interest in the joint model. We assume that the prior distributions for the unknown parameters are all independent, so that $\pi(\boldsymbol{\Theta}) = \pi(\beta_0)\pi(\upsilon_b)\pi(\Psi_b)\pi(\upsilon_w)\pi(\Psi_w)\prod_{i=1}^N \pi(\rho_i)\pi(\alpha_1, \alpha_2, \alpha_3)\pi(\boldsymbol{\gamma})$. Thus the joint distribution of the data and unknowns is

$$\begin{aligned} P(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\Theta}) \propto & \prod_{i=1}^N \frac{1}{\sqrt{(2\pi)^{n_i} \mid \boldsymbol{\Sigma}_i \mid}} \exp\Big(-\frac{1}{2}\left[\boldsymbol{X}_i - f(\boldsymbol{t}_i; \beta_0, \boldsymbol{b}_i)\right]^\top \\ & \boldsymbol{\Sigma}_i^{-1}\left[\boldsymbol{X}_i - f(\boldsymbol{t}_i; \beta_0, \boldsymbol{b}_i)\right]\Big) \\ & \times \prod_{l=1}^L \left\{\frac{1}{\sqrt{2\pi\sigma^2_{b_i}}} \exp\left(-\frac{b^2_{il}}{2\sigma^2_{b_i}}\right)\right\} \\ & \times \frac{1}{\sqrt{2\pi\Psi^2_b}} \exp\left\{-\frac{\left[\log\left(\sigma^2_{b_i}\right) - \upsilon_b\right]^2}{2\Psi^2_b}\right\} \\ & \times \frac{1}{\sqrt{2\pi\Psi^2_w}} \exp\left\{-\frac{\left[\log\left(\sigma^2_{w_i}\right) - \upsilon_w\right]^2}{2\Psi^2_w}\right\} \\ & \times \Phi\left(\eta\left(\sigma_{b_i}, \sigma_{w_i}, \boldsymbol{Z}_i\right)\right)^{Y_i} \\ & \times \left(1 - \Phi\left(\eta\left(\sigma_{b_i}, \sigma_{w_i}, \boldsymbol{Z}_i\right)\right)\right)^{1-Y_i} \times \pi(\boldsymbol{\Theta}) \end{aligned} \tag{6}$$

2.5 | Posterior Inference

Since we proposed several non-conjugate priors in our model, the posterior distribution is not available in closed form. Therefore, we implemented our joint model using **JAGS** and the **R2jags** package in R [32]. **JAGS** has several built-in sampling algorithms for non-conjugate distributions, including slice sampling, adaptive random walk Metropolis algorithm, and so forth. The main engine for **JAGS** is slice sampling, but **JAGS** can automatically choose samplers for updating parameter values. One benefit of choosing **JAGS** to implement our model is that the built-in glm module in **JAGS** is able to do block updating of β_0 and $\boldsymbol{b}_i$ as it recognizes the generalized linear mixed model structure in the longitudinal marker sub-model, which substantially increases the sampling efficiency.

3 | Simulation

Before applying our proposed model to any application settings, we conducted simulation studies. There are two goals associated with the simulation: (1) to test the performance of our

proposed model to see whether it has the capacity to recover the true parameter values in terms of bias in the posterior means, posterior variance, empirical coverage rates and average lengths of the nominal 95% credible intervals; (2) to compare with alternative two-stage approaches that can be used to obtain variances and then use variances as predictors of outcomes. Through our simulations, we demonstrated that our proposed joint model maintained satisfactory performance regardless of the magnitude of the measurement error in the longitudinal marker, whereas the commonly used two-stage approach failed to provide good results as the measurement error increased. Let θ_0 denote the true parameter value of θ and $\widehat{\theta}_r$ denote the posterior mean obtained from the rth simulation. We evaluated the performance of the simulations based on the following four criteria: (1) *Coverage Rate* by calculating $\frac{1}{R}\sum_{r=1}^{R} I\left(\theta_0 \in \left[L_r, U_r\right]\right)$ where L_r and U_r represents the 2.5% and 97.5% percentiles of the posterior draws from the rth simulation; (2) *Bias* by calculating $\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta_0\right)$; (3) *Average Interval Length* by calculating $\frac{1}{R}\sum_{r=1}^{R}\left(U_r - L_r\right)$; (4) *Root Mean Square Error (RMSE)* by calculating $\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\theta_0 - \widehat{\theta}_r\right)^2}$.

For the simulations, we ran three chains and for each chain we ran 8000 iterations with the first 4000 as burn-in. We assessed the convergence of the models by Gelman-Rubin $\widehat{R}$ measure. $\widehat{R}$ is approximately the square root of the ratio of the total variance of a given set of parameter draws to their within-chain variance; the $\widehat{R}$ values for all model parameters are smaller than 1.1, which is generally viewed as a satisfactory threshold [33].

3.1 | Simulation Settings

In the simulation, we generated $n_i = 600$ observations for each individual with $N = 150$ individuals in each simulation, and we conducted $R = 200$ replicates of the simulation. To mimic the heart rate data that would be modeled from the Michigan Work-Life Study in the Application section, we used the heart rate data in the Application section to obtain the true parameter values of the longitudinal marker sub-model that would be applied to our simulations. Specifically, we fit the longitudinal marker sub-model to 40 randomly selected participants' heart rate data from the Application data set and we used the exact same priors as described in the previous section. Note that we selected the time knots k_l's in the cubic B-splines to be every 2nd percentile of the observed time points. As we mentioned previously, ignoring the uncertainty in the first stage, which partially arises from the measurement errors in the longitudinal data, can lead to biased inference in two-stage approaches. Therefore, we conducted two sets of simulation comparisons with different magnitudes of measurement errors in the longitudinal marker by varying the prior distribution of σ_{w_i}'s. Specifically, in the low measurement error simulation setting, we drew σ_{w_i}'s from a log-normal distribution with a mean υ_w and variance Ψ_w^2, where υ_w and Ψ_w were set to the values obtained from fitting the application data mentioned above. In the high measurement error simulation setting, we drew σ_{w_i}'s from the exact same distribution but we increased υ_w to a larger value.

3.1.1 | Simulation Setting 1: Low Measurement Error in the Longitudinal Marker

We first simulated the longitudinal marker data for each individual independently based on the parameter values we obtained from the application data set:

$$\begin{aligned} &X_{ij} = \beta_0 + \sum_{l=1}^{L} b_{il}\xi_{l,3}(t_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, 600; \quad b_{il} \sim N\left(0, \sigma_{b_i}^2\right), \quad l = 1, \ldots, L \\ &\epsilon_{ij} = \rho_i \epsilon_{ij-1} + w_{ij}, \quad w_{ij} \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_{w_i}^2\right), \quad j = 1, \ldots, 600 \\ &\beta_0 = 81.083, \quad \log\left(\sigma_{b_i}^2\right) \sim N(5, 1^2), \quad \log\left(\sigma_{w_i}^2\right) \sim N\left(-\mathbf{3.8}, 0.75^2\right), \quad \rho_i = 0.998 \end{aligned} \tag{7}$$

A sample of simulated heart rate data of four randomly selected individuals in one simulation is shown in Figure 1. From the figure, we notice that the simulated values fall into a reasonable range of heart rate and the trend looks like normal heart rate data, which means that our proposed longitudinal marker sub-model can generate HR-like data from reasonable parameter values.

As the next step, we generated the binary outcomes based on σ_{b_i}'s and σ_{w_i}'s simulated from the longitudinal sub-model. We assumed that the outcome Y_i's came from a Bernoulli distribution with $P\left(Y_i = 1\right) = \Phi\left(\eta\left(\boldsymbol{S}_i\right)\right)$, where Φ is the CDF of distribution $N(0, 1)$ and set

$$\eta\left(\boldsymbol{S}_i\right) = \alpha_1 + \alpha_2 \times \log\left(\sigma_{w_i} \times 10\right) + \alpha_3 \times \log\left(\frac{\sigma_{b_i}}{10}\right) \tag{8}$$

We scaled σ_{b_i}'s and σ_{w_i}'s inside the log function to facilitate model convergence. We set the true values of $\boldsymbol{\alpha}$ to be $(0.3, -1, 0.5)$ so that the generated binary outcomes have a similar number of cases and controls.

3.1.2 | Simulation Setting 2: High Measurement Error in the Longitudinal Marker

In the second set of simulations, we chose a larger value for υ_w, the mean of the log-normal distribution from which σ_{w_i}'s were drawn. Other than that, all other parameter values in the longitudinal marker sub-model remained the same as in simulation setting 1:

$$\begin{aligned} &X_{ij} = \beta_0 + \sum_{l=1}^{L} b_{il}\xi_{l,3}(t_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, 600; \quad b_{il} \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_{b_i}^2\right), \quad l = 1, \ldots, L \\ &\epsilon_{ij} = \rho_i \epsilon_{ij-1} + w_{ij}, \quad w_{ij} \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_{w_i}^2\right), \quad j = 1, \ldots, 600 \\ &\beta_0 = 81.083, \quad \log\left(\sigma_{b_i}^2\right) \sim N(5, 1^2), \quad \log\left(\sigma_{w_i}^2\right) \sim N\left(\mathbf{1.8}, 0.75^2\right), \quad \rho_i = 0.998 \end{aligned} \tag{9}$$

Figure 2 displays the simulated heart rate data of four randomly selected individuals under the high measurement error setting. We notice that the simulated values and trends still resemble real heart rate data except that there are more fluctuations and the overall trends appear to be less smooth than the simulated data under the low measurement error setting.

As the next step, we generated the binary outcomes based on the two types of variances simulated from the longitudinal sub-model. We used the same procedure for drawing the outcome Y_i's; however, we adjusted how we scaled σ_{b_i}'s and σ_{w_i}'s inside the log function to facilitate model convergence and we assigned new values to the coefficients $\boldsymbol{\alpha}$ to ensure the generated binary

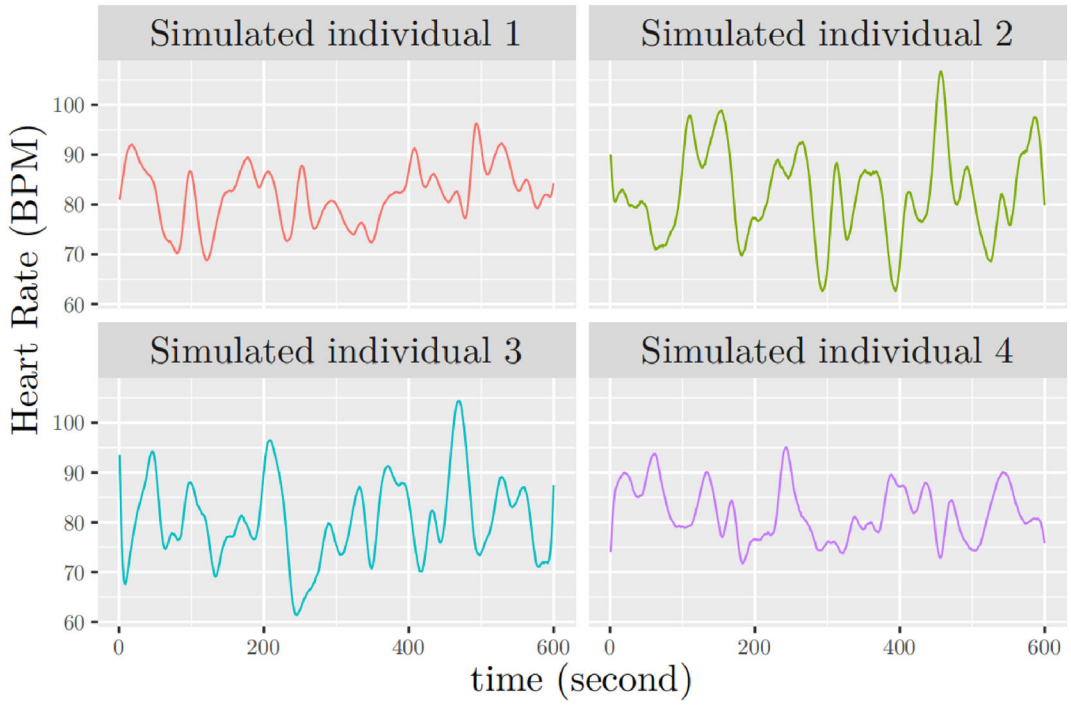


FIGURE 1 | Simulated heart rate data of four randomly selected individuals in simulation setting 1.

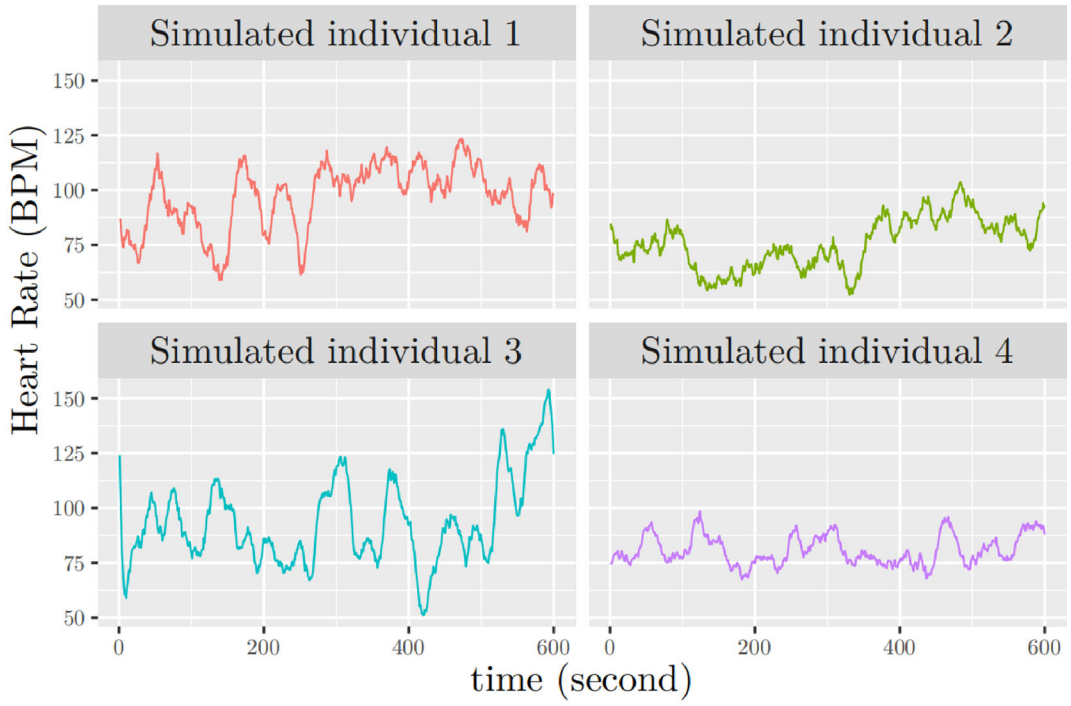


FIGURE 2 | Simulated heart rate data of four randomly selected individuals in simulation setting 2.

outcomes were balanced. Specifically, we set

$$\eta(\boldsymbol{S}_i) = \alpha_1 + \alpha_2 \times \log(\sigma_{w_i}) + \alpha_3 \times \log\left(\frac{\sigma_{b_i}}{10}\right) \tag{10}$$

and we set the true values of $\boldsymbol{\alpha}$ to be $(-0.6, 1, -1.5)$, again retaining an approximately equal number of positive and negative outcomes.

3.2 | Alternative Methods

To compare with our joint model, we used a two-stage modeling strategy as an alternative method. In the first stage, we fit a linear mixed effects model for the simulated heart rate data across all simulated participants using lme() function from the **nlme** package in R:

TABLE 1 | Simulation results including coverage rate, bias, average interval length, and RMSE of $\boldsymbol{\alpha}$ comparing joint model and two-stage approach under simulation setting 1 when the measurement error in the longitudinal marker is low.

True values	Model	Coverage (%)	Average interval length	Bias	RMSE
$\alpha_1 = 0.3$	Joint model	96.5	0.65	0.01	0.15
	Two-stage model	94.0	0.63	−0.05	0.17
$\alpha_2 = -1$	Joint model	95.0	1.20	−0.02	0.28
	Two-stage model	93.0	1.22	0.01	0.34
$\alpha_3 = 0.5$	Joint model	94.0	0.91	0.01	0.21
	Two-stage model	92.0	0.87	0.03	0.24

TABLE 2 | Simulation results including coverage rate, bias, average interval length, and RMSE of $\boldsymbol{\alpha}$ comparing joint model and two-stage model approach under simulation setting 2 when the measurement error in the longitudinal marker is high.

True values	Model	Coverage (%)	Average interval length	Bias	RMSE
$\alpha_1 = -0.6$	Joint model	94.5	1.28	−0.04	0.34
	Two-stage model	11.5	1.54	−1.17	1.23
$\alpha_2 = 1$	Joint model	95.5	1.38	0.07	0.37
	Two-stage model	24.0	1.64	1.01	1.09
$\alpha_3 = -1.5$	Joint model	92.5	1.37	−0.14	0.41
	Two-stage model	61.0	1.73	−0.73	0.85

$$X_{ij} = \beta_0 + \sum_{l=1}^{L} b_{il}\xi_{l,3}(t_{ij}) + \epsilon_{ij} \tag{11}$$

When fitting the linear mixed effects model, we assumed that the residuals had a constant variance across time and followed a first-order autoregressive (AR1) structure.

To obtain the subject-specific long-term variability $\hat{\sigma}^2_{b_i}$'s, for each subject i, we calculated the empirical variance of its estimated random effects: $\hat{\sigma}^2_{b_i} = \frac{1}{L-1}\sum_{l=1}^{L}\left(\hat{b}_{il} - \bar{\hat{b}}_i\right)^2$, $\bar{\hat{b}}_i = \frac{1}{L}\sum_{l=1}^{L}\hat{b}_{il}$. To approximate the subject-specific short-term variability $\hat{\sigma}_{w_i}$'s, for each subject i, we first calculated the fitted values $\hat{X}_{ij} = \hat{\beta}_0 + \sum_{l=1}^{L}\hat{b}_{il}\xi_{l,3}(t_{ij}) + \hat{\rho}\hat{\epsilon}_{ij-1}$ to get the estimated "white noise" $\hat{w}_{ij} = X_{ij} - \hat{X}_{ij}$ and then computed $\hat{\sigma}^2_{w_i} = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(\hat{w}_{ij} - \bar{\hat{w}}_i\right)^2$, $\bar{\hat{w}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{w}_{ij}$.

In the second stage, we regressed the simulated outcome on the scaled long-term variability $\log\left(\frac{\hat{\sigma}_{bi}}{10}\right)$ and short-term variability $\log(\hat{\sigma}_{w_i} \times 10)$ or $\log(\hat{\sigma}_{w_i})$ using a logistic regression with a probit link function by glm() function in R.

3.3 | Simulation Results

Table 1 shows the results of comparing the two-stage approach with our proposed joint model when the measurement error in the longitudinal marker is low. For our proposed joint model, we observe that the coverage rates of the true parameter values in the outcome sub-model are all greater than or equal to 94% and the estimated coefficients are nearly unbiased. The two-stage model has a slightly larger bias and somewhat lower coverage in this low-variance setting.

Table 2 presents the results of comparing the two approaches when the measurement error in the longitudinal marker is high. We observe that our proposed joint model still retained satisfactory bias and coverage rates of parameters in the outcome sub-model, although the interval lengths are wider than in the low variance setting. However, the two-stage approach now suffers from severe bias and low coverage rates due to the measurement error introduced through use of empirical estimates in the two-stage modeling.

The main takeaway from the two simulation comparisons is that while the two-stage approach can yield comparable results to our proposed joint model when the measurement error in the longitudinal marker is low, the joint model is superior over the two-stage approach in handling high measurement error scenarios and producing more valid inference. These findings again confirm what we learned from the existing literature and emphasize the advantages of joint modeling. Overall, the above results demonstrate that our proposed model is capable of recovering the true parameter values with high coverage rates and low bias.

4 | Application to the Michigan Work-Life Study

4.1 | Study Description

The Michigan Work-Life Study was a randomized controlled study designed to investigate anticipatory and perseverative stress associated with potentially discriminatory and prejudicial

situations [34]. Two hundred and thirty-two healthy women from a medium-sized metropolitan statistical area in Michigan, who identified as Black or African American and were between the ages of 18 and 30 years participated in a 2-h laboratory social stress test. After completing a 40-min online survey that included sociodemographic, psychosocial, personality, and health questions, participants were randomized to one of four experimental groups for the laboratory social stress test. Throughout the entirety of the laboratory test, continuous heart rate and periodic blood pressure and stress perception data and saliva samples were collected. After exclusion of participants with substantial missing continuous heart rate data, our final analytic sample was 165. We focus our analyses on heart rate as a well-studied marker of the stress response system that is typically analyzed using information averaged over periods of time. This study was approved by the University of Michigan Institutional Review Board (HUM00134109).

The laboratory social stress test was designed around the TSST, a well-documented and well-validated mode of eliciting a biological and psychological stress response [35]. In the TSST, a participant is asked to deliver a 5-min talk about some aspect of themselves and a 5-min math task in front of an audience. The audience members are generally instructed to sit and listen to the talk and math task without facial or bodily expression. (Studies have shown that when audience members nod their heads, smile, or provide other positive cues, the TSST does not elicit the stress response from the participant [36].) In this specific study, there were three experimental groups and one control group. The three experimental groups differed in the racial composition and the racial attitudes of the audience. In Groups 1 and 2, the audience was composed of four White women and one White woman study staff member. In Group 3, the audience was composed of four Black women and one Black woman study staff member. The control group (Group 4) did not include an audience to determine whether the audience-based TSST elicited an appropriate stress response, but included a Black woman staff member. In Group 1, the participants were led to believe that the audience members were part of a human resources/administrative organization that held racially-hostile views. In Groups 2 and 3, the participants were led to believe that the audience members were part of an organization that held racially-inclusive views. Table 3 shows a detailed group specification.

In studies using the TSST, participants have been exposed to the audience in different ways. Some studies convene a small, live audience for the participant. The limitations of this approach include the lab scheduling for hundreds of participants and, perhaps more importantly, the risk of unintentional positive cues from the audience. Others present an opaque panel and inform the participant that an audience is hidden behind the panel and in reality, there is no audience. The limitation to this approach is that it may not elicit the stress response because the facial and bodily cues are not visible to the participant. Therefore, this study simulated a virtual conference call on a large-screen television. The audience included a recording of four women appearing to listen to a talk. The advantages to this approach are that facial and bodily cues were well-controlled and that, within their experimental groups, all study participants were exposed to the exact same women and facial and bodily cues. In fact, those in Groups 1 and 2 were exposed to the exact same audience, but were led to believe that they held different racial attitudes.

TABLE 3 | Group specification.

Group	Audience
Group 1 (WH)	Four White women and one White woman study staff member with racially-hostile views
Group 2 (WF)	Four White women and one White woman study staff member with racially-inclusive views
Group 3 (BF)	Four Black women and one Black woman study staff member with racially-inclusive views
Group 4 (C)	Control group including a Black woman study staff member

Note: Group 1 to Group 3 are the three experimental groups and Group 4 is the control group. WH stands for "White hostile," WF stands for "White friendly," BF stands for "Black friendly," and C stands for "control."

Upon the arrival to the laboratory, participants completed the informed consent process. Notably, participants were led to believe that this study was focused on the ways in which potential employers view different job applicants. Only at the end of the study, during a full debriefing, were participants informed of the true research objective of the study. After obtaining informed consent, staff connected participants to an Omron blood pressure cuff (Omron Healthcare, Kyoto, Japan) and an Empatica E4 wristband (Empatica, Cambridge, MA).

The entire social stress test period is illustrated in Figure 3. The TSST proceeded in timed sections as follows, which are separated by delineations we call "Marks." In the first 10-min period (ANTICIPATION, bracketed by Mark 1 and Mark 2), the staff member explained that the participant would be giving a talk to a virtual audience of members from the (fictional) "Michigan Human Resources Administrators Organization" and in this 5-min talk, they were to discuss why they were the ideal candidate for their dream job, whatever that job may be. The staff then walked them through a (fictional) recent survey completed by the members of this organization. The survey results showed that the organization was either composed of almost entirely White members (Groups 1 and 2) or very diverse with respect to racial composition (Group 3). Further, the survey results showed that members supported racially-hostile views (Group 1) or racially-inclusive views (Groups 2 and 3). In the remaining minutes of this period, participants could prepare for their talk with notes, but were instructed that they could not use these notes during their talk. During the 10-min STRESS period, the staff revealed a large screen television with a supposed live audience virtual meeting of four women. During the first 5 min, bracketed by Mark 3 and Mark 4, the staff member instructed the participant to explain to the audience why they are the ideal candidate for their dream job. If they finished speaking in less than 5 min, they were instructed to continue speaking. During the second 5 min, bracketed by Mark 4 and Mark 5, the staff member instructed the participant to sequentially subtract 13, starting from 1022. If they made a mistake, they were instructed to

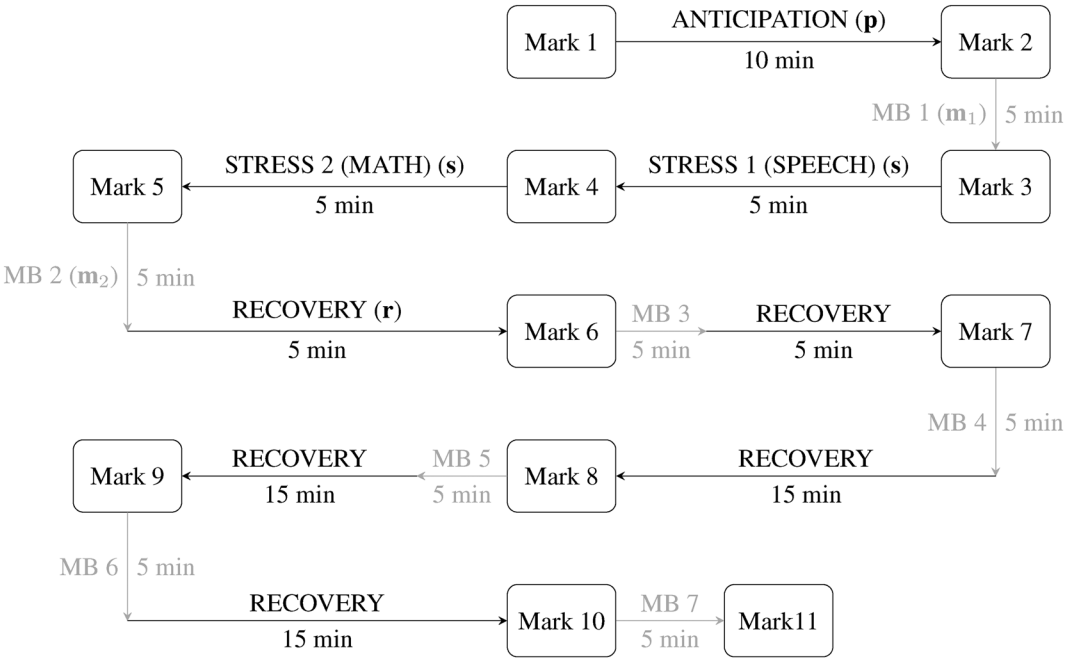


FIGURE 3 | Flowchart of the entire experimental period. The focus of the statistical analysis is on the experimental period from Mark 1 to Mark 6. We labeled the five periods from Mark 1 to Mark 6 by $\mathbf{p}$, $\mathbf{m}_1$, $\mathbf{s}$, $\mathbf{m}_2$, and $\mathbf{r}$, respectively.

begin again. After this stress period, the staff member turned off the television and instructed the participant to sit quietly for the remaining roughly 90 min (RECOVERY, bracketed by Mark 5 through Mark 11). We limit our analysis to heart rate measurements between Mark 1 and Mark 6, to focus on the anticipation and stress periods.

In between the ANTICIPATION period and the first RECOVERY period, the Empatica E4 collected continuous heart rate information; at specific time points (called a "Measure Break" or "MB"), staff collected blood pressure, stress perception, and saliva. For a variety of reasons, these MB periods may have provided ambiguous responses; for example, some participants may have continued to think about the previous period while others may have been sufficiently distracted by the blood pressure and saliva collection so that they are not thinking about the previous period. Hence, we do include these MB periods in our analysis; however, these MB periods are not the focus of our application.

4.2 | Model Framework

4.2.1 | Longitudinal Marker Sub-Model

The longitudinal marker of interest in this application is heart rate over time, which was collected each second continuously throughout the entire experimental period. There were an average of 2153 heart rate observations per woman between Mark 1 and Mark 6; the number of observations varied from a minimum of 1878 to a maximum of 3872. Figure 4 shows the heart rate measurements from Mark 1 to Mark 6 for eight randomly selected participants, with two participants randomly selected from each of the four experimental groups. Table 4 presents the summary statistics of the average and variability of heart rate measurements over time across four experimental groups and five experimental periods.

We fit the cubic B-spline model with AR1 as shown in the method section. We selected the time knots k_l's to be every 2nd percentiles of the observation time t_{ij}. However, we assumed that the long-term variability and the short-term variability differ by experimental periods (e.g., anticipation, stress) rather than the uniform long-term and short-term variability in (1) and (5). We included five periods bracketed by Mark 1 through Mark 6 (see Figure 3): the 10-min anticipation period, denoted by $\mathbf{p}$; the first (nuisance) 5-min measurement break, denoted by $\mathbf{m}_1$; the 10-min stress period, denoted by $\mathbf{s}$; the second (nuisance) 5-min measurement break, denoted by $\mathbf{m}_2$; and a 5-min recovery period, denoted by $\mathbf{r}$. Our long-term variability model is given by

$$\begin{aligned} b_{il} \overset{\text{i.i.d}}{\sim} N\big(0, \sigma^{(p)2}_{b_i} \times I\big(\text{period}_{il} = \boldsymbol{p}\big) + \sigma^{(\boldsymbol{m_1})2}_{b_i} \times I\big(\text{period}_{il} = \boldsymbol{m_1}\big) \\ + \sigma^{(s)2}_{b_i} \times I\big(\text{period}_{il} = \boldsymbol{s}\big) + \sigma^{(\boldsymbol{m_2})2}_{b_i} \times I\big(\text{period}_{il} = \boldsymbol{m_2}\big) + \sigma^{(r)2}_{b_i} \\ \times I\big(\text{period}_{il} = \boldsymbol{r}\big)\big), l = 1, \ldots, L \end{aligned} \tag{12}$$

where, $I(\cdot)$ is an indicator function indicating whether the lth random effect of the ith participant falls into a specific experimental period. We also assumed that the long-term variability in these different periods had their own hyperpriors:

$$\begin{aligned} &\log\Big(\sigma^{(p)2}_{b_i}\Big) \sim N\Big(\upsilon^{(p)}_b, \Psi^{(p)2}_b\Big), \quad \log\Big(\sigma^{(\boldsymbol{m_1})2}_{b_i}\Big) \sim N\Big(\upsilon^{(\boldsymbol{m_1})}_b, \Psi^{(\boldsymbol{m_1})2}_b\Big) \\ &\log\Big(\sigma^{(s)2}_{b_i}\Big) \sim N\Big(\upsilon^{(s)}_b, \Psi^{(s)2}_b\Big), \quad \log\Big(\sigma^{(\boldsymbol{m_2})2}_{b_i}\Big) \sim N\Big(\upsilon^{(\boldsymbol{m_2})}_b, \Psi^{(\boldsymbol{m_2})2}_b\Big), \\ &\log\Big(\sigma^{(r)2}_{b_i}\Big) \sim N\Big(\upsilon^{(r)}_b, \Psi^{(r)2}_b\Big) \end{aligned} \tag{13}$$

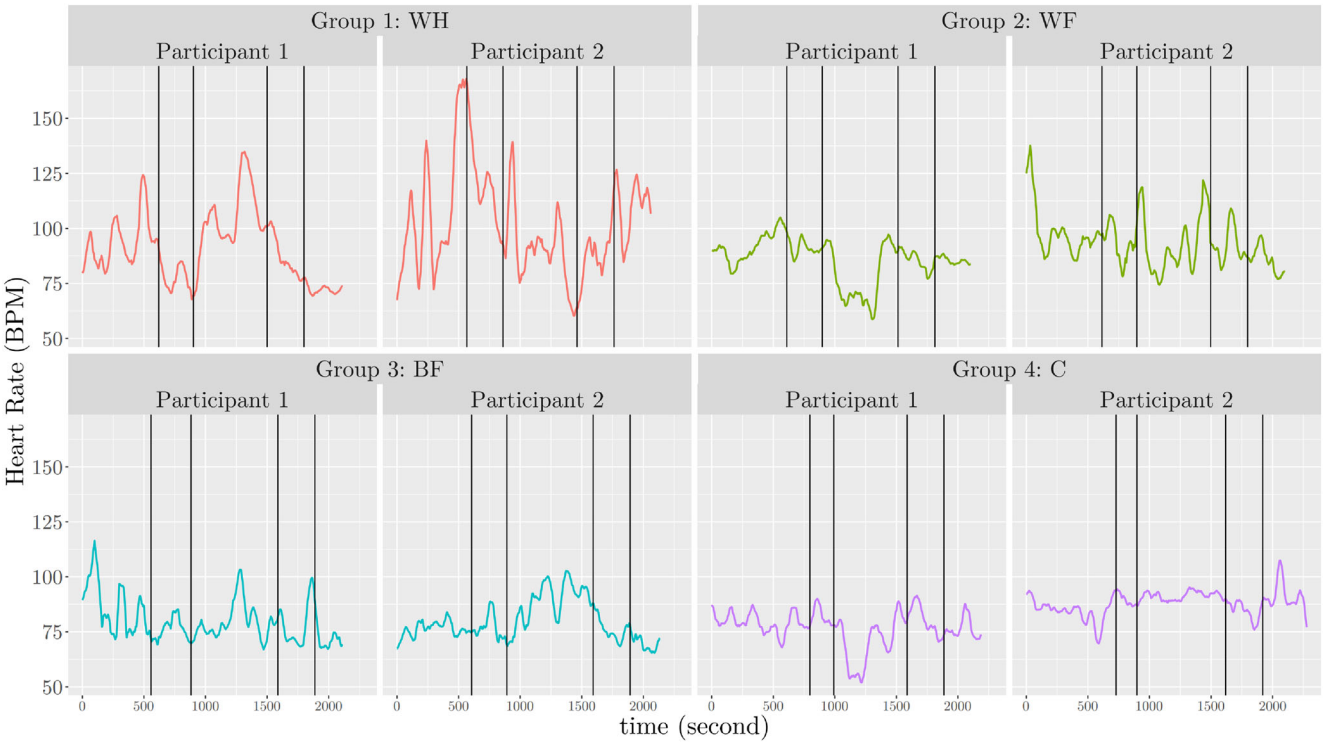


FIGURE 4 | Heart rate measurements from Mark 1 to Mark 6 for eight randomly selected participants, with two participants randomly selected from each of the four experimental groups. The four solid lines correspond to Mark 2 representing the end of the anticipation period and the start of the first measure break, Mark 3 representing the start of the stress period, Mark 4 representing the end of the stress period and the start of the second measure break, and Mark 5 representing the start of the recovery period, respectively. BF, Black friendly audience group; C, control group; WF, White friendly audience group; WH, White hostile audience group.

TABLE 4 | Summary statistics mean (SD) of the mean and standard deviation of heart rate measurements over time across four experimental groups and five experimental periods.

	WH		WF		BF		C	
	HR: mean	**HR: SD**	**HR: mean**	**HR: SD**	**HR: mean**	**HR: SD**	**HR: mean**	**HR: SD**
Anticipation	83.07 (8.69)	9.21 (6.75)	84.77 (16.18)	7.94 (5.15)	82.08 (9.24)	8.37 (6.09)	80.80 (6.33)	9.93 (8.10)
MB1	82.46 (10.59)	8.43 (6.03)	82.88 (10.84)	7.60 (4.85)	81.73 (10.06)	6.90 (3.95)	86.35 (16.14)	9.97 (9.35)
Stress	85.53 (10.92)	11.25 (5.88)	87.59 (13.09)	10.93 (5.60)	83.97 (8.79)	11.64 (5.04)	85.99 (13.44)	12.46 (6.81)
MB2	80.53 (10.68)	7.61 (4.46)	81.20 (11.76)	8.86 (6.50)	78.02 (7.68)	8.15 (3.83)	80.41 (13.44)	8.28 (5.23)
Recovery	77.65 (8.99)	7.32 (5.92)	77.62 (10.42)	7.68 (5.29)	76.56 (10.02)	6.89 (5.57)	73.78 (7.98)	6.13 (3.21)

Abbreviations: BF, Black friendly audience group; C, control group; MB, measure break; SD, standard deviation; WF, White friendly audience group; WH, White hostile audience group.

Similarly, we extended (6) to

$$w_{ij} \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_{w_i}^{(p)2} \times I\left(\text{period}_{ij} = \boldsymbol{p}\right) + \sigma_{w_i}^{(\boldsymbol{m_1})2} \times I\left(\text{period}_{ij} = \boldsymbol{m_1}\right) + \sigma_{w_i}^{(s)2} \times I\left(\text{period}_{ij} = \boldsymbol{s}\right) + \sigma_{w_i}^{(\boldsymbol{m_2})2} \times I\left(\text{period}_{ij} = \boldsymbol{m_2}\right) + \sigma_{w_i}^{(r)2} \times I\left(\text{period}_{ij} = \boldsymbol{r}\right)\right), j = 1, \ldots, n_i \tag{14}$$

Similarly, as the long-term variability, we assumed that the short-term variability in these different periods had their own hyperpriors:

$$\log\left(\sigma_{w_i}^{(p)2}\right) \sim N\left(\upsilon_w^{(p)}, \Psi_w^{(p)2}\right), \quad \log\left(\sigma_{w_i}^{(\boldsymbol{m_1})2}\right) \sim N\left(\upsilon_w^{(\boldsymbol{m_1})}, \Psi_w^{(\boldsymbol{m_1})2}\right),$$
$$\log\left(\sigma_{w_i}^{(s)2}\right) \sim N\left(\upsilon_w^{(s)}, \Psi_w^{(s)2}\right), \quad \log\left(\sigma_{w_i}^{(\boldsymbol{m_2})2}\right) \sim N\left(\upsilon_w^{(\boldsymbol{m_2})}, \Psi_w^{(\boldsymbol{m_2})2}\right),$$
$$\log\left(\sigma_{w_i}^{(r)2}\right) \sim N\left(\upsilon_w^{(r)}, \Psi_w^{(r)2}\right) \tag{15}$$

The posterior means of the square root of individual-level variances, $\sigma_{w_i}^{(\cdot)}$'s and $\sigma_{b_i}^{(\cdot)}$'s, across the four experimental groups in each of the experimental period is shown in Figure 5.

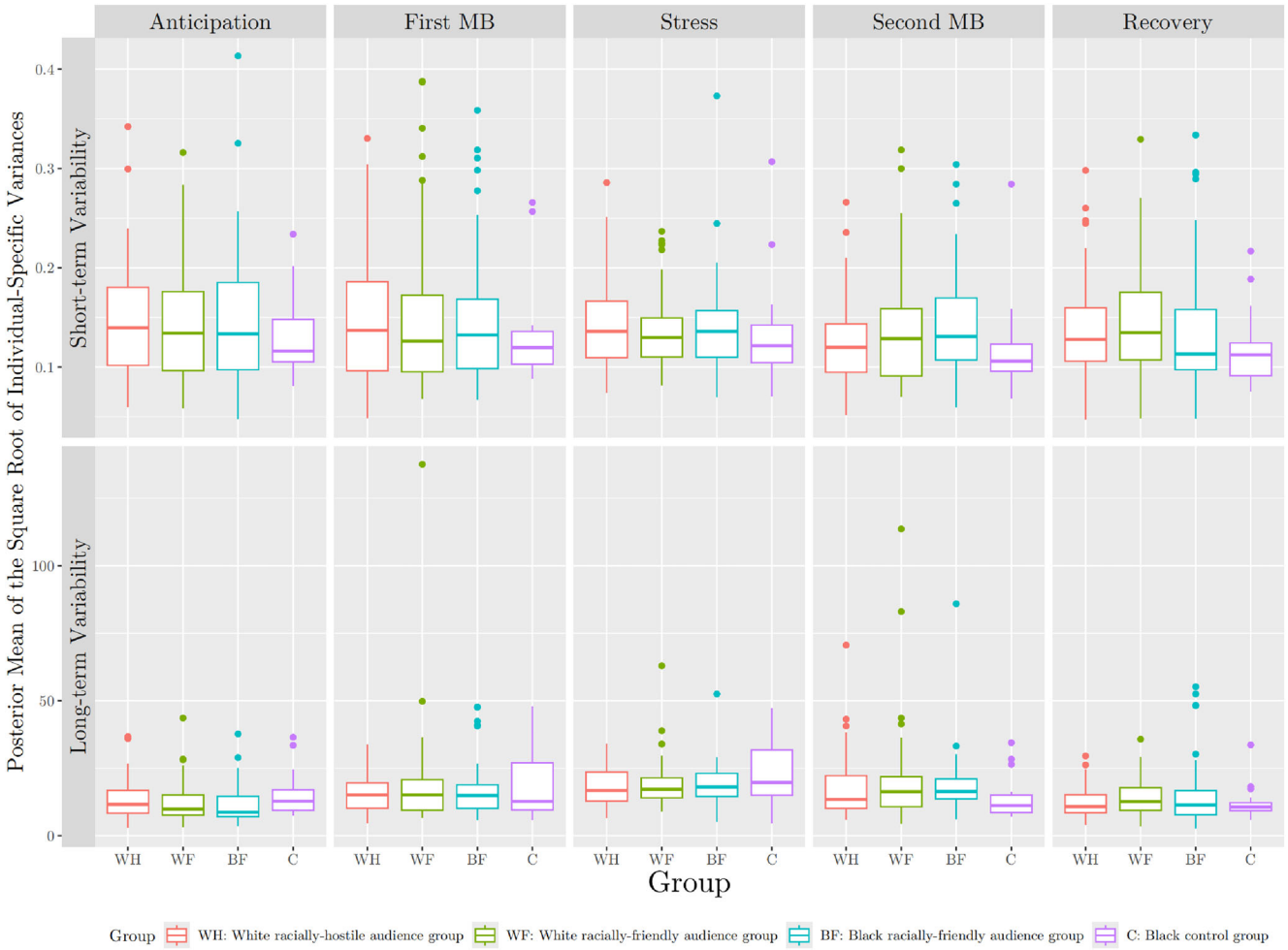


FIGURE 5 | Short- and long-term variability in each of the experimental periods across groups characterized by the posterior means of the square root of individual-specific variances. The top five panels correspond to the posterior means of $\sigma_{w_i}^{(p)}$s, $\sigma_{w_i}^{(m_1)}$s, $\sigma_{w_i}^{(s)}$s, $\sigma_{w_i}^{(m_2)}$s, and $\sigma_{w_i}^{(r)}$s, respectively and the bottom five panels correspond to the posterior means of $\sigma_{b_i}^{(p)}$s, $\sigma_{b_i}^{(m_2)}$s, $\sigma_{b_i}^{(s)}$s, $\sigma_{b_i}^{(m_2)}$s, and $\sigma_{b_i}^{(r)}$s, respectively.

4.2.2 | Outcome Sub-Model

The outcome of interest in this application is the binary group assignment: Group 1 versus Group 2 with Group 1 ($Y_i = 0$) as the reference group, Group 2 versus Group 3 with Group 2 ($Y_i = 0$) as the reference group, Group 1 versus Group 4 with Group 4 ($Y_i = 0$) as the reference group, Group 2 versus Group 4 with Group 4 ($Y_i = 0$) as the reference group, or Group 3 versus Group 4 with Group 4 ($Y_i = 0$) as the reference group. Pairwise group comparisons follow the substantive meaning represented by the experimental conditions. Comparing Groups 1 and 2 allows us to focus on potential variation introduced by the racial attitudes of the audience (hostile or friendly) because the racial composition of the audience was the same between groups. Comparing Groups 2 and 3 allows us to focus on the racial composition (Black or White) because the racial attitudes of the audience were the same. We also compared each group with the control group in order to examine whether there was an impact of public speaking (the core of the TSST). Note that we used group assignment as our outcome mainly to study its association with short-term/long-term variability in heart rate, which is thought to reflect the anticipation and perseveration about potential discriminatory and prejudicial situations reflected by the group assignment, rather than testing the causal relationship. Let $\boldsymbol{S}_i = \left[\sigma_{w_i}^{(p)}, \sigma_{w_i}^{(m_1)}, \sigma_{w_i}^{(s)}, \sigma_{w_i}^{(m_2)}, \sigma_{w_i}^{(r)}, \sigma_{b_i}^{(p)}, \sigma_{b_i}^{(m_1)}, \sigma_{b_i}^{(s)}, \sigma_{b_i}^{(m_2)}, \sigma_{b_i}^{(r)}\right]^\top$ be a vector of individual-specific short-term and long-term variabilities across different experimental periods, we extended the linear prediction function in (2) to

$$\begin{aligned}\eta(\boldsymbol{S}_i) = \alpha_1 &+ \alpha_2 \times \log\left(\sigma_{w_i}^{(p)} \times 10\right) + \alpha_3 \times \log\left(\sigma_{w_i}^{(m_1)} \times 10\right) \\ &+ \alpha_4 \times \log\left(\sigma_{w_i}^{(s)} \times 10\right) \\ &+ \alpha_5 \times \log\left(\sigma_{w_i}^{(m_2)} \times 10\right) + \alpha_6 \times \log\left(\sigma_{w_i}^{(r)} \times 10\right) \\ &+ \alpha_7 \times \log\left(\frac{\sigma_{b_i}^{(p)}}{10}\right) + \alpha_8 \times \log\left(\frac{\sigma_{b_i}^{(m_1)}}{10}\right) + \alpha_9 \times \log\left(\frac{\sigma_{b_i}^{(s)}}{10}\right) \\ &+ \alpha_{10} \times \log\left(\frac{\sigma_{b_i}^{(m_2)}}{10}\right) + \alpha_{11} \times \log\left(\frac{\sigma_{b_i}^{(r)}}{10}\right)\end{aligned} \tag{16}$$

TABLE 5 | Hypothesized differences in stress response across TSST periods and the experimental groups.

TSST period	WF vs. WH	BF vs. WF	WH vs. C	WF vs. C	BF vs. C
Anticipation	WH > WF	ND	WH > C	WF > C	BF > C
Stress	WH > WF	BF > WF	WH > C	WF > C	BF > C
Recovery	WH > WF	BF > WF	WH > C	WF > C	BF > C

Abbreviations: BF, Black friendly audience group; C, control group; ND, no difference; WF, White friendly audience group; WH, White hostile audience group.

where we scaled the long-term and short-term variances inside the log function to make the chains in the posterior sampling procedure have better mixing performance. We did not include any other covariates in the outcome model since the participants were randomly assigned and their covariates values were balanced. The parameters we are interested in are the regression coefficients in the linear prediction function $\alpha_1, \ldots, \alpha_{11}$.

For the analysis, we ran three chains with 18 000 iterations each, treating the first 9000 as burn-in. We again assessed the convergence of the models by Gelman-Rubin $\hat{R}$ measure, showing maximum values under 1.1 and thus good convergence. A Dell computer with Intel Xeon E5-2650 v4 @ 2.20 GHz CPU (24 cores) and 256 GB memory was used to perform the analysis of the application dataset and it took an average of 165.7 h and 22.23 GB to run one application joint model.

4.3 | Results

Drawing from the TSST and stress biology literatures as well as the humanities and social science literatures, we hypothesize the following patterns of results across the TSST periods and experimental groups (see Table 5). We hypothesize that, during all periods, participants exposed to the White, racially-hostile audience will experience more stress compared to those exposed to the White racially-friendly audience. We further hypothesize that, during the anticipation period, participants exposed to the Black, racially-friendly audience and the White, racially-friendly audience will experience roughly similar levels of stress. This is because they will have been told the same things about the speaking activity and the audience. The only difference between these groups is the race of the staff member. However, both the Black and the White staff members interacted in a civil but stoic manner with participants. Interacting with another Black woman who is not overtly welcoming in a new space (the lab) may be a source of stress. The social science literature discuss the importance of social cues of affirmation among Black women, even among strangers, in public. Thus, during the stress and recovery periods, we hypothesize that participants exposed to the stoic audience of Black women, even though they were primed to believe that they were racially-friendly, will experience greater stress compared to those exposed White, racially-friendly audience. Finally, we hypothesize that participants exposed to any of the audiences will experience greater stress compared to those who were exposed only to the (Black) staff member. However, there is the possibility that those exposed to the White, friendly audience may experience lower levels of stress compared to those in the control group, because this latter group still had to interact with the Black staff member who carried herself in a stoic manner.

The results from the application of our analytic approach suggest that there is a difference in the stress response meaning of second-to-second (short-term) and minute-to-minute (long-term) patterns in variability. Table 6 provides the results from all of the separate models together to facilitate interpretation. Note that each column represents the results from a separate model. When considering the results for second-to-second variability, participants in the control group (conceptualized as the lowest stress exposure group, as there was no exposure to any audience) did not exhibit higher variability in heart rate in any phase compared to any of the other experimental groups (Table 6, top panel). However, when considering minute-to-minute variability, participants in the control group exhibited greater variability in heart rate compared to each of the three other exposure groups at different phases (Table 6, bottom panel). For example, participants in the control group exhibited greater minute-to-minute variability compared to those exposed to the Black friendly audience during the anticipation and stress phases. While this pattern was not universal, as we discuss below, the difference in the pattern of results for the second-to-second and minute-to-minute variability suggests that each of these types of variability captures something different about the stress response process.

Minute-to-minute (long-term) variability may capture what is thought to be captured by standard HRV that is averaged over time. The general literature on stress and HRV suggests that greater mean HRV, averaged over time (e.g., a 5-min period), is related to lower heart rate and lower acute stress exposure. Compared to participants in each of the three audience-based stress exposure groups, those in the control group (conceptualized as the least stressful exposure group) exhibited greater variability in heart rate. Specifically, participants in the control group exhibited greater variability compared to those exposed to the White hostile audience during the break period that immediately followed the anticipation period (α_8: −23.62; 95% CI: −36.20, −13.73); compared to those exposed to the White friendly audience during the breaks immediately following both the anticipation and stress periods (α_8: −14.52; 95% CI: −23.13, −6.38; α_{10}: −5.87; 95% CI: −11.91, −0.79); and compared to those exposed to the Black friendly audience during the anticipation and stress periods (α_7: −15.01; 95% CI: −25.68, −2.01; α_9: −8.64; 95% CI: −17.22, −0.43). This suggests that those in the control group experienced less stress exposure compared to those in the other three groups.

Notably, however, this pattern of greater minute-to-minute variability with the control group compared to other groups was not universal. If we are to understand the minute-to-minute variability as inversely related to stress response, then, during the stress period, participants exposed to the White, racially-friendly

TABLE 6 | Coefficients and 95% credible intervals for each of the experimental group comparisons for short-term variability ($\alpha_2, \ldots, \alpha_6$) and long-term variability ($\alpha_7, \ldots, \alpha_{11}$).

Coefficient (period)	WF vs. WH (ref, $Y_i = 0$)	BF vs. WF (ref, $Y_i = 0$)	WH vs. C (ref, $Y_i = 0$)	WF vs. C (ref, $Y_i = 0$)	BF vs. C (ref, $Y_i = 0$)
α_1 (Intercept)	−0.12	0.02	4.83	0.31	0.80
	(−0.73, 0.45)	(−0.55, 0.60)	(0.69, 9.99)	(−3.44, 4.60)	(−2.83, 4.78)
α_2 (Anticipation)	0.19	0.22	4.59	8.18	**12.50**
	(−1.12, 1.55)	(−0.83, 0.59)	(−6.25, 16.12)	(−3.00, 19.79)	**(2.03, 22.96)**
α_3 (MB1)	0.16	−0.11	8.67	2.82	−4.23
	(−0.84, 1.17)	(−0.96, 0.74)	(−0.04, 18.82)	(−4.27, 10.38)	(−12.16, 3.01)
α_4 (Stress)	−1.35	−0.12	6.73	0.11	2.14
	(−2.91, 0.10)	(−1.54, 1.25)	(−2.81, 17.56)	(−8.91, 9.36)	(−5.80, 10.89)
α_5 (MB2)	0.60	0.64	−2.07	**10.87**	**8.65**
	(−0.54, 1.79)	(−0.33, 1.60)	(−12.09, 7.88)	**(2.27, 20.72)**	**(1.02, 18.12)**
α_6 (Recovery)	0.14	−0.71	5.98	**11.42**	−3.66
	(−0.90, 1.24)	(−1.83, 0.34)	(−3.28, 16.57)	**(1.58, 22.01)**	(−12.95, 5.52)
α_7 (Anticipation)	−0.72	−0.08	5.88	−4.13	**−15.01**
	(−1.87, 0.28)	(−0.86, 0.72)	(−2.15, 14.06)	(−11.80, 2.24)	**(−25.68, −2.01)**
α_8 (MB1)	0.30	−0.17	**−23.62**	**−14.52**	2.97
	(−0.66, 1.37)	(−0.85, 0.48)	**(−36.20, −13.73)**	**(−23.13, −6.38)**	(−3.24, 9.41)
α_9 (Stress)	0.53	0.08	1.91	**12.30**	**−8.64**
	(−0.43, 1.54)	(−0.82, 1.01)	(−4.65, 9.28)	**(5.69, 20.13)**	**(−17.22, −0.43)**
α_{10} (MB2)	0.07	−0.05	4.32	**−5.87**	**12.50**
	(−0.56, 0.71)	(−0.65, 0.56)	(−1.79, 12.05)	**(−11.91, −0.79)**	**(0.90, 24.19)**
α_{11} (Recovery)	0.25	0.25	−1.76	−3.98	−1.44
	(−0.54, 1.07)	(−0.36, 0.94)	(−11.64, 7.32)	(−12.41, 4.75)	(−8.52, 5.58)

Abbreviations: BF, Black friendly audience group; C, control group; MB, measure break; ref, reference group; WF, White friendly audience group; WH, White hostile audience group.
Note: Bold indicates if the 95% credible interval does not cross and cover 0.

audience experienced less stress compared to those in the control group. As discussed above, it may be that interaction with a stoic, Black staff member resulted in similar stress experiences as delivering a talk in front of a stoic, White, racially-friendly audience. Our results also suggest that, during the break immediately following the stress period, participants exposed to the Black, racially-friendly audience experienced less stress compared to those in the control group. It may be that there are similar levels of stress experienced when interacting with a stoic panel of Black women virtually and with a Black staff member in person.

Our results suggest that second-to-second (short-term) variability in heart rate captures something different about the stress response process and is perhaps positively (rather than inversely) correlated with the stress experience. During the stress period, participants exposed to the White, racially-hostile audience experienced greater variability compared to those exposed to the White, racially-friendly audience (α_4: −1.35; 95% CI: −2.91, 0.10), although the 95% CI included zero. Similarly, when there were differences in the variability in heart rate between any of the three audience-based groups and the control group, our results suggest that participants delivering a talk in front of an audience experienced greater stress. This pattern is consistent with that shown with minute-to-minute variability.

4.4 | Posterior Predictive Model Checking

To assess whether the proposed model produces measurements that approximate the true data from the Michigan Work-Life Study, we further conducted a "goodness-of-fit" test on both the longitudinal marker sub-model and the binary outcome sub-model based on the posterior predictive distribution (PPD). Let $\boldsymbol{X}_i^{rep}$ be the vector of replicated longitudinal data for subject i. We calculated the posterior predictive p-values [37] across all subjects, which examined the probability that a test statistic calculated from the replicated data $\boldsymbol{X}_i^{rep}$ is more extreme than the same test statistic calculated from the observed data $\boldsymbol{X}_i$. For the longitudinal marker data, the replicated data $\boldsymbol{X}_i^{rep}$ is drawn from the PPD $f\left(\boldsymbol{X}_i^{rep}|\boldsymbol{X}_i\right) = \int f\left(\boldsymbol{X}_i^{rep}|\boldsymbol{\Theta}, \boldsymbol{X}_i\right) p\left(\boldsymbol{\Theta}|\boldsymbol{X}_i\right) d\boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is the vector of unknown model parameters. We obtained $\boldsymbol{X}_i^{rep}$ from the following steps, note that we included the first-order autocorrelation structure when simulating $\boldsymbol{X}_i^{rep}$:

1. We calculated the fitted values of the jth observation for the ith subject from the mean structure of the longitudinal marker $f\left(t_{ij}; \beta_0, \boldsymbol{b}_i\right) = \beta_0 + \sum_{l=1}^{L} b_{il}\xi_{l,3}\left(t_{ij}\right)$.
2. We generated "white noise" w_{ij} from $N\left(0, \sigma_{w_i}^{(p)2} \times I\left(\text{period}_{ij} = \boldsymbol{p}\right) + \sigma_{w_i}^{(\boldsymbol{m_1})2} \times I\left(\text{period}_{ij} = \boldsymbol{m_1}\right) + \sigma_{w_i}^{(s)2} \times$

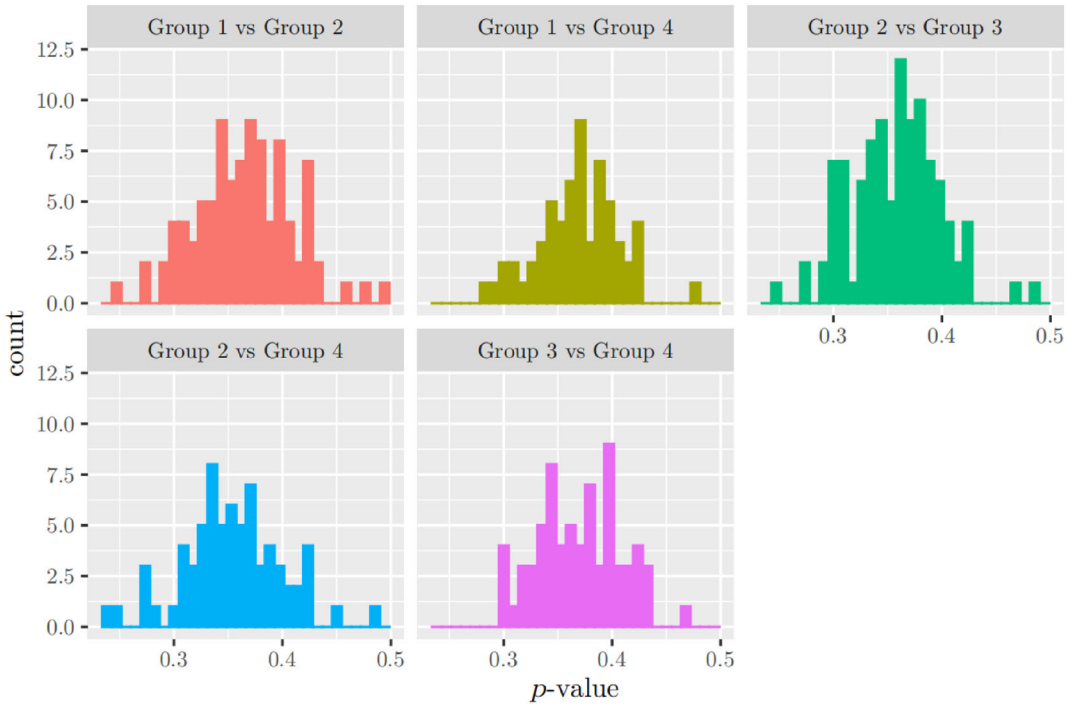


FIGURE 6 | Posterior predictive check for longitudinal marker data across all five models in the application. Each panel corresponds to the histogram of the posterior predictive *p*-values across all individuals in one specific model. The top three panels correspond to the model comparing Group 1 with Group 2 (median: 0.364), the model comparing Group 1 with Group 4 (median: 0.372) and the model comparing Group 2 with Group 3 (median: 0.362), respectively; the bottom two panels correspond to the model comparing Group 2 with Group 4 (median: 0.353) and the model comparing Group 3 with Group 4 (median: 0.368), respectively.

$$I\left(\text{period}_{ij}=\boldsymbol{s}\right)+\sigma^{(\boldsymbol{m_2})2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{m_2}\right)+\sigma^{(\boldsymbol{r})2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{r}\right)\Big).$$

3. We generated the initial error ϵ_{i0} by $\epsilon_{i0}\sim N\left(0,\frac{\sigma^{(p)2}_{w_i}}{1-\rho_i^2}\right)$ and then generated a series of errors with AR1 correlation structure as $\epsilon_{ij}=\rho_i\epsilon_{ij-1}+w_{ij}$.
4. We calculated fitted values incorporating AR1 correlation structure by $\tilde{f}\left(t_{ij};\beta_0,\boldsymbol{b}_i\right)=f\left(t_{ij};\beta_0,\boldsymbol{b}_i\right)+\rho_i\epsilon_{ij-1}$ with w_{ij} left over as random errors.
5. We drew the jth element of $\boldsymbol{X}_i^{rep}$ from the PPD as $X_{ij}^{rep}=f\left(t_{ij};\beta_0,\boldsymbol{b}_i\right)+\epsilon_{ij}$, $j=1,\ldots,n_i$.

Then, we defined a chi-square discrepancy statistics as $T_i\left(\boldsymbol{X}_i;\beta_0,\boldsymbol{b}_i,\boldsymbol{\sigma}^2_{w_i}\right)=\sum_{j=1}^{n_i}\frac{\left(X_{ij}-\tilde{f}\left(t_{ij};\beta_0,\boldsymbol{b}_i\right)\right)^2}{\sigma^2_{w_{ij}}}$, where $\boldsymbol{\sigma}^2_{w_i}$ is a vector of short-term variability across time with the jth element $\sigma^2_{w_{ij}}=\sigma^{(p)2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{p}\right)+\sigma^{(\boldsymbol{m_1})2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{m_1}\right)+\sigma^{(s)2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{s}\right)+\sigma^{(\boldsymbol{m_2})2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{m_2}\right)+\sigma^{(r)2}_{w_i}\times I\left(\text{period}_{ij}=\boldsymbol{r}\right)$. We computed 500 values of $T_i\left(\boldsymbol{X}_i;\beta_0,\boldsymbol{b}_i,\boldsymbol{\sigma}^2_{w_i}\right)$ and $T_i\left(\boldsymbol{X}_i^{rep};\beta_0,\boldsymbol{b}_i,\boldsymbol{\sigma}^2_{w_i}\right)$ from the observed longitudinal marker data and 500 draws from the posterior of β_0, $\boldsymbol{b}_i$ and $\boldsymbol{\sigma}^2_{w_i}$. We compared the test statistics obtained from the observed data and the replicated data by the posterior predictive *p*-value defined as $P\left(T_i\left(\boldsymbol{X}_i;\beta_0,\boldsymbol{b}_i,\boldsymbol{\sigma}^2_{w_i}\right)<T_i\left(\boldsymbol{X}_i^{rep};\beta_0,\boldsymbol{b}_i,\boldsymbol{\sigma}^2_{w_i}\right)|\boldsymbol{X}_i\right)$. Since we fit five joint models in the application, we conducted the posterior predictive check on the longitudinal marker on all five models. Figure 6 displays the histograms of the posterior predictive *p*-values in each of the five models. Most of the computed *p*-values fall between 0.2 and 0.5 and the median values of the posterior predictive *p*-values in these models are all around 0.36, which indicates that the replicated values from the longitudinal marker sub-model are relatively close to the observed values used to fit the model and the longitudinal sub-model fits the observed marker data well.

We further checked the PPD of the binary outcome Y_i. For the binary outcome data, the replicated data Y_i^{rep} is drawn from a Bernoulli distribution with probability $\pi_i=\Phi\left(\eta\left(\boldsymbol{\sigma}^2_{w_i},\boldsymbol{\sigma}^2_{b_i},\boldsymbol{\alpha}\right)\right)$, where $\boldsymbol{\sigma}^2_{w_i}$ denotes the vector of short-term variability across time, $\boldsymbol{\sigma}^2_{b_i}$ denotes the vector of long-term variability across time and $\boldsymbol{\alpha}$ represents the vector of regression coefficients of the outcome sub-model. We considered the test statistic as the total number of events $T(\boldsymbol{Y})=\sum_{i=1}^{N}Y_i$ and we computed 500 values of $T\left(\boldsymbol{Y}^{rep}\right)$ from 500 draws from the posterior of $\boldsymbol{\sigma}^2_{w_i}$, $\boldsymbol{\sigma}^2_{b_i}$ and $\boldsymbol{\alpha}$. We compared the test statistics obtained from the replicated data with the total number of events in the observed binary outcomes by the posterior predictive *p*-values $P\left(T(\boldsymbol{Y})<T\left(\boldsymbol{Y}^{rep}\right)|\boldsymbol{Y}\right)$. Similarly to what we did with checking the longitudinal marker sub-model, we repeated the posterior predictive check on the binary outcomes in all five models. Table 7 shows the total number of events in the observed binary outcomes $T(\boldsymbol{Y})$ and the posterior predictive *p*-value of each of the five models. The *p*-values are all around 0.4, which means that the replicated binary outcomes from the outcome sub-model are close to the observed data and the outcome sub-model is a good fit of the observed data.

TABLE 7 | Posterior predictive check for the binary outcomes across all five models in the application.

	$T(Y)$	p
Group 1 vs. Group 2	52	0.45
Group 1 vs. Group 4	43	0.36
Group 2 vs. Group 3	52	0.46
Group 2 vs. Group 4	52	0.38
Group 3 vs. Group 4	52	0.34

Note: $T(\boldsymbol{Y})$ represents the total number of events in the observed binary outcomes from each specific model.

5 | Discussion

This paper presents a joint modeling approach that connects variabilities in the longitudinal marker data with a binary outcome. Through simulation studies, we have demonstrated that our joint model performs better than the commonly used two-stage approach when relating longitudinal measurements with a cross-sectional outcome. We applied our method to the Michigan Work-Life Study and investigated the association between social stress and variabilities in the longitudinal measurements of heart rate. We found that participants exposed to a higher vigilance situation typically exhibited a significantly higher short-term variability and lower long-term variability when they were asked to prepare for or perform the TSST.

Our work is important in terms of both methodological development of joint modeling and application of heart rate data. Joint modeling proves to be superior to the two-stage approach in that joint modeling yields much more accurate estimates. Unlike the two-stage approach, where uncertainty from the first stage cannot be carried over to the second stage, joint modeling allows for the propagation of the uncertainty from the longitudinal marker sub-model to the outcome sub-model. As shown in the simulation studies, our proposed joint model consistently achieves high coverage rates and low bias for the unknown parameters whereas the two-stage approach performed poorly when the measurement error in the longitudinal marker data increases. Our work extended the existing joint modeling approaches to intensive longitudinal measurements setting which consists of thousands of measurements and incorporated both the residual and random effects variabilities as predictors in the outcome sub-model, which is a novel contribution compared to the previous joint modeling study on variabilities as predictors which focused primarily on the residual variances. We termed the variability of the random effects as "long-term" variability and the variability of the residuals as "short-term" variability, which characterized individual-level variabilities into two types. This approach of joint modeling is rare in the literature and could be useful in modeling intensive longitudinal marker data, especially when the number of random effects in the modeling is high and individual interpretation of each random effect becomes challenging. Our work is also innovative in terms of the application of heart rate data. The traditional approach of dealing with longitudinal measurements of heart rate in existing studies is through HRV, where the time and frequency domain of HRV are first calculated and then used in subsequent analysis. In contrast, our model is able to jointly model longitudinal heart rate and outcomes as well as connecting different levels of variabilities in heart rate with health outcomes.

We acknowledge certain limitations in our work. First of all, we suffered from a small sample size when applying our proposed model to the actual data. There were only 18 participants in the control group and approximately 50 participants in each experimental group after removing participants with a significant amount of missing heart rate data, which makes models comparing each experimental group with the control group converge in a much longer time, as we were using a small number of data points to estimate eleven coefficients. We also made a strong assumption of the correlations among heart rate measurements over time by specifying an AR1 autocorrelation structure. The large replication of measurements and the small time differences between adjacent measurements can make the AR1 autocorrelation structure questionable and relaxing the autocorrelation structure to AR2 or higher orders may achieve a better fit to the data, although in this application we were unable to get stable estimates from such a model. Another limitation is the considerable time and resources required to fit the proposed method. Exploring strategies to improve the efficiency of the fitting process can help make the method more accessible. A minor consideration is that instead of fitting five separate probit models for pairwise comparisons, we can also try fitting a multinomial probit model as an alternative.

Several extensions of this work can be made in the future. First, although our model checking suggested reasonable fit to the assumed distributions, alternative distributional assumptions could be considered: for example, Gamma distributions for the prior on the variances, or mixture distributions for the heart rate itself. Another extension would be to include additional biomarkers in the analysis. For example, in the Michigan Work-Life Study, the Empatica E4 wristband not only collects continuous heart rate measurements but also monitors continuous electrodermal activity (EDA) measurements and skin temperature. Introducing more biomarkers into the analysis can help us find out how variabilities in different biological signals are connected with stress. However, this extension requires modeling the correlation between biomarkers, demanding more efficient modeling techniques given that modeling one single intensive longitudinal marker is already time-consuming. Another potential extension of the current work could be to allow individual-specific variances to vary by time. In our application, we implemented a simple version of this concept, where we allowed individual-specific variances to differ across experimental periods. A more sophisticated approach, however, is to model σ_{e_i}'s as a function of time t_{ij}. This extension may better capture the changes in biomarkers since such data like heart rate are highly variable.

Author Contributions

Mingyan Yu, Zhenke Wu, and Michael R. Elliott designed the statistical methods and Mingyan Yu implemented the methods. Margaret Hicken designed and conducted the Michigan Work-Life Study and Mingyan Yu applied the methods to the study. Mingyan Yu, Zhenke Wu, Margaret Hicken, and Michael R. Elliott wrote and reviewed the manuscript.

Acknowledgments

This work was supported in part by funds from the National Institutes of Health (NIH grants R56AG-066693 and R01MD-016244).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. R. Henderson, P. Diggle, and A. Dobson, "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics* 1, no. 4 (2000): 465–480.

2. L. Wu, X. J. Hu, and H. Wu, "Joint Inference for Nonlinear Mixed-Effects Models and Time to Event at the Presence of Missing Data," *Biostatistics* 9, no. 2 (2008): 308–320.

3. J. Horrocks and V. M. J. Den Heuvel, "Prediction of Pregnancy: A Joint Model for Longitudinal and Binary Data," *Bayesian Analysis* 4 (2009): 523–538.

4. S. Shiffman, A. Stone, and M. Hufford, "Ecological Momentary Assessment," *Annual Review of Clinical Psychology* 4 (2008): 1–32.

5. R. Schuster, R. Mermelstein, and D. Hedeker, "Ecological Momentary Assessment of Working Memory Under Conditions of Simultaneous Marijuana and Tobacco Use," *Addiction* 111 (2016): 1466–1476.

6. X. Li and D. Hedeker, "A Three-Level Mixed-Effects Location Scale Model With an Application to Ecological Momentary Assessment Data," *Statistics in Medicine* 31 (2012): 3192–3210.

7. J. Oleson, M. Jones, E. Jorgensen, and Y. Wu, "Statistical Considerations for Analyzing Ecological Momentary Assessment Data," *Journal of Speech, Language, and Hearing Research* 65 (2022): 344–360.

8. S. Shiffman, S. Scholl, J. Mao, S. Ferguson, D. Hedeker, and H. Tindle, "Ecological Momentary Assessment of Temptations and Lapses in Non-Daily Smokers," *Psychopharmacology* 237 (2020): 2353–2365.

9. D. McNeish, D. Mackinnon, L. Marsch, and R. Poldrack, "Measurement in Intensive Longitudinal Data," *Structural Equation Modeling: A Multidisciplinary Journal* 28 (2021): 807–822.

10. R. J. Carroll, "Variances Are Not Always Nuisance Parameters," *Biometrics* 59, no. 2 (2003): 211–220.

11. J. Palmier-Claus, P. J. Taylor, P. Gooding, G. Dunn, and S. Lewis, "Affective Variability Predicts Suicidal Ideation in Individuals at Ultra-High Risk of Developing Psychosis: An Experience Sampling Study," *British Journal of Clinical Psychology* 51, no. 1 (2012): 72–83.

12. Y. Fang, D. B. Forger, E. Frank, S. Sen, and C. Goldstein, "Day-to-Day Variability in Sleep Parameters and Depression Risk: A Prospective Cohort Study of Training Physicians," *npj Digital Medicine* 4, no. 1 (2021): 28.

13. B. J. Taylor, K. A. Matthews, B. P. Hasler, et al., "Bedtime Variability and Metabolic Health in Midlife Women: The SWAN Sleep Study," *Sleep* 39, no. 2 (2016): 457–465.

14. F. Gao, J. P. Miller, C. Xiong, J. A. Beiser, M. Gordon, and Group OHTSO, "A Joint-Modeling Approach to assess the Impact of Biomarker Variability on the Risk of Developing Clinical Outcome," *Statistical Methods and Applications* 20 (2011): 83–100.

15. M. R. Elliott, M. D. Sammel, and J. Faul, "Associations Between Variability of Risk Factors and Health Outcomes in Longitudinal Studies," *Statistics in Medicine* 31, no. 23 (2012): 2745–2756.

16. B. Jiang, M. R. Elliott, M. D. Sammel, and N. Wang, "Joint Modeling of Cross-Sectional Health Outcomes and Longitudinal Predictors via Mixtures of Means and Variances," *Biometrics* 71, no. 2 (2015): 487–497.

17. D. Hedeker, R. Mermelstein, and H. Demirtas, "Modeling Between-Subject and Within-Subject Variances in Ecological Momentary Assessment Data Using Mixed-Effects Location Scale Models," *Statistics in Medicine* 31 (2012): 3328–3336.

18. M. Kikuya, T. Ohkubo, H. Metoki, et al., "Day-by-Day Variability of Blood Pressure and Heart Rate at Home as a Novel Predictor of Prognosis: The Ohasama Study," *Hypertension* 52 (2008): 1045–1050.

19. P. F. Hsu, H. M. Cheng, C. H. Wu, et al., "High Short-Term Blood Pressure Variability Predicts Long-Term Cardiovascular Mortality in Untreated Hypertensives but Not in Normotensives," *American Journal of Hypertension* 29 (2016): 806–813.

20. C. K. Endukuru and S. Tripathi, "Evaluation of Cardiac Responses to Stress in Healthy Individuals—A Non Invasive Evaluation by Heart Rate Variability and Stroop Test," *International Journal of Science and Research* 5 (2016): 286–289.

21. H. Du and L. Wang, "Reliabilities of Intraindividual Variability Indicators With Autocorrelated Longitudinal Data: Implications for Longitudinal Study Designs," *Multivariate Behavioral Research* 53 (2018): 502–520.

22. S. Wang, T. H. McCormick, and J. T. Leek, "Methods for Correcting Inference Based on Outcomes Predicted by Machine Learning," *National Academy of Sciences of the United States of America* 117, no. 48 (2020): 30266–30275.

23. W. Ye, X. Lin, and J. M. Taylor, "Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data—A Two-Stage Regression Calibration Approach," *Biometrics* 64, no. 4 (2008): 1238–1246.

24. L. Wu, W. Liu, G. Y. Yi, and Y. Huang, "Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues," *Journal of Probability and Statistics* 2012 (2012): 640153.

25. M. C. Wu and R. J. Carroll, "Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process," *Biometrics* 44 (1988): 175–188.

26. J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data," *Journal of Clinical Oncology* 28, no. 16 (2010): 2796–2801.

27. M. Alsefri, M. Sudell, M. García-Fiñana, and R. Kolamunnage-Dona, "Bayesian Joint Modelling of Longitudinal and Time to Event Data: A Methodological Review," *BMC Medical Research Methodology* 20 (2020): 1–17.

28. F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health* 5 (2017): 258.

29. A. Gelman, "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on article by Browne and Draper)," *Bayesian Analysis* 1 (2006): 515–534.

30. A. S. Conlon, J. M. Taylor, and M. R. Elliott, "Surrogacy Asessment Using Principal Stratification When Surrogate and Outcome Measures Are Multivariate Normal," *Biostatistics* 15 (2014): 266–283.

31. R. P. Ghosh, B. Mallick, and M. Pourahmadi, "Bayesian Estimation of Correlation Matrices of Longitudinal Data," *Bayesian Analysis* 16, no. 3 (2021): 1039–1058, https://doi.org/10.1214/20-BA1237.

32. Y. S. Su and M. Yajima, *R2jags: Using R to Run 'JAGS'. R Package Version 0.7-1* (2021).

33. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (Boca Raton, FL: CRC Press, 2013).

34. P. J. Sawyer, B. Major, B. J. Casad, S. S. Townsend, and W. B. Mendes, "Discrimination and the Stress Response: Psychological and Physiological Consequences of Anticipating Prejudice in Interethnic Interactions," *American Journal of Public Health* 102, no. 5 (2012): 1020–1026.

35. C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'—A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting," *Neuropsychobiology* 28, no. 1-2 (1993): 76–81.

36. U. S. Wiemers, D. Schoofs, and O. T. Wolf, "A Friendly Version of the Trier Social Stress Test Does Not Activate the HPA Axis in Healthy Men and Women," *Stress* 16, no. 2 (2013): 254–260.

37. A. Gelman, X. L. Meng, and H. Stern, "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica* 6 (1996): 733–760.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.