Supplementary Materials

Ilaria Colpizzi[1], Federico Calà[2], Lorenzo Frassineti[2], Antonio Lanatà[2], Claudia Manfredi[2], Claudio Sica[3], Igor Marchetti[3], & Corrado Caudek[4]

[1] Department of Life Sciences, University of Trieste, Trieste, Italy

[2] Department of Information Engineering, University of Florence, Florence, Italy

[3] Department of Health Sciences, University of Florence, Florence, Italy

[4] Department of NEUROFARBA, University of Florence, Florence, Italy

Author Note

Correspondence concerning this article should be addressed to Ilaria Colpizzi, Trieste, Italy. E-mail: ilaria.colpizzi@units.it

Supplementary Materials

Reliability of Personality Measures

Overview

Reliability was evaluated for both the full baseline PID-5 questionnaire (220 items) and the brief EMA-based PID-5 assessment (15 items, 3 per domain). For the EMA measures, which involve repeated assessments nested within persons, we employed multilevel reliability estimation following Lai (2021), which distinguishes between-person reliability (consistency of person-level means) from within-person reliability (consistency of occasion-level deviations).

Baseline PID-5 Questionnaire

The full PID-5 was administered at baseline (T1) using the standard 220-item version. Reliability was computed using Cronbach's alpha and McDonald's omega for each domain and for the total score. Results are reported for both the full sample and a cleaned sample excluding participants who failed attention check items (catch trials embedded at positions 68 and 161 in the questionnaire).

Results. All domains showed acceptable to excellent internal consistency (Table S1). Psychoticism exhibited the highest reliability ($\alpha = .87$, $\omega = .90$), consistent with its larger item pool (33 items). Antagonism showed the lowest reliability ($\alpha = .67$, $\omega = .82$), though McDonald's omega—which accounts for item heterogeneity—indicated adequate composite reliability. The total PID-5 score demonstrated excellent reliability ($\alpha = .96$, $\omega = .98$). Excluding participants who failed attention checks produced slightly lower but comparable estimates, indicating that careless responding had minimal impact on scale properties in this sample.

Table S1

Internal consistency of baseline PID-5 domains.

Domain	Items	α (Full)	ω (Full)	α (Clean)	ω (Clean)
Negative Affectivity	23	0.783	0.851	0.767	0.843
Detachment	22	0.754	0.844	0.741	0.840
Antagonism	21	0.670	0.819	0.627	0.796
Disinhibition	22	0.747	0.843	0.725	0.824
Psychoticism	33	0.870	0.898	0.864	0.892
Total PID-5	220	0.965	0.979	0.964	0.977

EMA-Based Brief PID-5

The brief PID-5 administered via EMA comprised 15 items (3 per domain), selected based on factor loadings and domain representativeness from prior validation work (Bottesi et al., 2024). Because these items were assessed repeatedly across approximately 20 occasions per participant, standard single-level reliability estimates are inappropriate. Instead, we employed the multilevel composite reliability framework described by Lai (2021), which partitions variance into within-person and between-person components and computes separate reliability indices for each level.

Multilevel Reliability Framework. Following Lai (2021), we estimated three reliability indices:

- α_{2L} **(Two-level alpha)**: Overall reliability of observed scores, pooling within- and between-person variance. Relevant when scores are used without distinguishing levels.

- α_{B} **(Between-person alpha)**: Reliability of person-level means (aggregated across occasions). This is the relevant index when EMA scores are averaged to create a single trait estimate per person, as in our moderation analyses.

- α_W **(Within-person alpha)**: Reliability of occasion-level deviations from each person's mean. Relevant for detecting momentary fluctuations in personality states.

The same decomposition was applied using McDonald's omega (ω_{2L}, ω_B, ω_W), which relaxes the assumption of tau-equivalence.

Results. The multilevel reliability analysis revealed a clear pattern (Table S2):

1. **Between-person reliability was high** ($\alpha_B = .87$, $\omega_B = .85$). This indicates that person-level trait estimates derived from aggregating across EMA occasions are measured with good precision. This is the critical index for our moderation analyses, which use person-level latent trait scores as predictors.

2. **Within-person reliability was adequate** ($\alpha_W = .73$, $\omega_W = .72$). This suggests that the brief scale can detect meaningful occasion-to-occasion fluctuations in personality states, though with more measurement error than at the between-person level. This is expected given the brevity of the scale (15 items total).

3. **Two-level reliability was good** ($\alpha_{2L} = .83$, $\omega_{2L} = .81$), reflecting the overall quality of observed scores when levels are pooled.

Table S2

Multilevel reliability of EMA-based brief PID-5 (15 items).

Reliability Index	Estimate	95% CI Lower	95% CI Upper
α_{2L} (two-level)	0.826	0.802	0.845
α_B (between)	0.866	0.835	0.889
α_W (within)	0.727	0.687	0.758
ω_{2L} (two-level)	0.814	—	—
ω_B (between)	0.852	—	—
ω_W (within)	0.724	—	—

Interpretation for the Present Study. The high between-person reliability (α_B = .87) supports the validity of using EMA-derived personality scores as person-level moderators of vocal stress responses. Because our moderation model estimates latent trait scores from the repeated EMA observations (see Statistical Models section), measurement error is explicitly modeled rather than ignored. The multilevel measurement model in our Stan implementation can be viewed as formalizing the reliability structure documented here: the latent trait θ_{id} represents the "true" person-level standing on domain d, while the occasion-specific observations X_{nd} are treated as noisy indicators with residual variance σ_d^{ema} capturing within-person fluctuation and measurement error.

The adequate within-person reliability (α_W = .73) also suggests that the brief EMA measure could support analyses of state-level personality-voice covariation, though such analyses would require denser sampling than the present design to achieve adequate statistical power.

Descriptive Statistics

Descriptive statistics are reported for $N = 119$ female participants. Participants contributed on average 27 EMA assessments ($SD = 4.16$, range = 12–31).

Design and Coverage

Table S3 summarizes the study design, including the number of participants, voice observations, and EMA assessments. All 119 participants contributed usable voice observations in each assessment period (Baseline, Pre-exam, Post-exam), consistent with the balanced, three-phase design.

Table S3

Study design and data coverage.

N	EMA M	EMA SD	EMA Range	Voice Obs. (per period)
119	27	4.16	12–31	119

Voice Outcomes by Assessment Period

Tables S4 and S5 report descriptive statistics for the two primary voice outcomes—fundamental frequency (F0) and normalized noise energy (NNE)—across the three assessment periods.

Across periods, mean F0 increased from Baseline (191.1 Hz) to Pre-exam (194.8 Hz) and decreased slightly at Post-exam (192.6 Hz), consistent with anticipatory stress-related elevation in pitch followed by partial recovery. NNE showed more negative values (indicating reduced glottal noise) from Baseline to Pre-exam, with partial return toward Baseline in the Post-exam period.

Table S5 summarizes within-person variability across the three assessment periods, reporting the distribution of participant-specific standard deviations.

Table S4

Voice outcome descriptives by assessment period.

Period	n	F0 M (Hz)	F0 SD	NNE M (dB)	NNE SD
Baseline	119	191.11	21.30	-26.43	2.63
Pre-exam	119	194.81	21.36	-27.12	3.25
Post-exam	119	192.63	23.12	-26.93	2.87

EMA Personality Domain Descriptives

EMA-based PID-5 domain descriptives are reported separately for between-person variability (Table S6) and within-person variability (Table S7).

Between-person descriptives indicate substantial inter-individual variability in EMA trait levels, while within-person descriptives confirm meaningful intra-individual fluctuation across repeated assessments. This pattern is consistent with the multilevel measurement model used to estimate latent trait scores and propagate measurement uncertainty into moderation effects.

Table S5

Within-person variability in voice parameters across periods.

Variable	n	Mean SD	SD of SDs	Min SD	Max SD
F0 within-person SD (Hz)	119	7.44	5.47	0	28.76
NNE within-person SD (dB)	119	1.73	1.11	0	5.58

Statistical Models

Main Effects of Exam-Related Stress on Vocal Parameters

Research Question. Before examining personality moderation, we first establish whether exam-related stress produces reliable changes in vocal acoustics. This model addresses the foundational question: Do vocal parameters (F0, NNE) change systematically across the three assessment phases—baseline, pre-exam (anticipatory stress), and post-exam (recovery)?

Model Specification. The model is a standard Bayesian hierarchical (multilevel) linear regression with repeated measures nested within participants. Stress effects are parameterized using two orthogonal contrasts:

- **Stress contrast (c_1)**: Compares pre-exam to baseline, capturing anticipatory stress-induced change.
- **Recovery contrast (c_2)**: Compares post-exam to pre-exam, capturing post-stressor trajectory.

Let y_n denote the vocal outcome for observation n, with participant index $s[n]$. The model specifies:

$$y_n \sim \text{Normal}(\mu_n, \sigma_y)$$

Table S6

EMA PID-5 descriptives: between-person variability (person means across occasions).

Domain	n	M	SD	Min	Max
Negative Affectivity	119	4.53	1.62	0.71	8.08
Detachment	119	2.04	1.62	0.00	6.19
Antagonism	119	0.87	1.17	0.00	5.29
Disinhibition	119	2.46	1.29	0.10	5.80
Psychoticism	119	1.25	1.44	0.00	5.73

where the linear predictor includes both fixed and random effects:

$$\mu_n = \alpha + u_{0,s} + (\beta_1 + u_{1,s}) \cdot c_{1n} + (\beta_2 + u_{2,s}) \cdot c_{2n}$$

The parameters are:

- α: Grand intercept (population-average baseline vocal level).
- β_1: Population-average stress effect (pre-exam vs. baseline).
- β_2: Population-average recovery effect (post-exam vs. pre-exam).
- $u_{0,s}$: Participant-specific random intercept.
- $u_{1,s}$: Participant-specific random slope for stress.
- $u_{2,s}$: Participant-specific random slope for recovery.
- σ_y: Residual standard deviation.

The random effects capture individual differences in baseline vocal characteristics (u_0), stress reactivity (u_1), and recovery patterns (u_2). A non-centered parameterization is used for computational efficiency:

Table S7

EMA PID-5 descriptives: within-person variability (participant SDs across occasions).

Domain	n	M of SD	SD of SDs	Min SD	Max SD
Negative Affectivity	119	1.32	0.42	0.57	2.42
Detachment	119	1.20	0.62	0.00	3.04
Antagonism	119	0.71	0.65	0.00	2.84
Disinhibition	119	1.17	0.45	0.30	2.31
Psychoticism	119	0.93	0.68	0.00	2.93

$$u_{k,s} = \tau_k \cdot z_{k,s}, \quad z_{k,s} \sim \text{Normal}(0, 1)$$

where τ_k are the random effect standard deviations.

Prior Specification. Priors were chosen to be weakly informative, incorporating domain knowledge about plausible parameter ranges while allowing the data to dominate inference (Table S8).

Interpretation. The key parameters of interest are β_1 (stress effect) and β_2 (recovery effect):

- A positive β_1 indicates that F0 increases from baseline to pre-exam, consistent with heightened autonomic arousal elevating vocal pitch.
- A negative β_2 would indicate that F0 decreases from pre-exam to post-exam, suggesting recovery toward baseline levels.

The random effect standard deviations (τ_1, τ_2, τ_3) quantify the degree of individual differences in baseline levels, stress reactivity, and recovery patterns. Large values of τ_2 or τ_3 would indicate substantial heterogeneity in how participants respond to stress—heterogeneity that might be explained by personality traits, motivating the

Table S8

Prior specifications for the main effects model.

Parameter	Prior	Rationale
α	Normal(220, 30)	Centered on typical female F0 (Hz)
β_1, β_2	Normal(0, 10)	Weakly informative; allows effects up to $\pm$20 Hz
τ_k	Exponential(0.5)	Weakly informative for random effect SDs
σ_y	Exponential(0.1)	Allows residual SD in plausible range

moderation analyses presented in the next section.

The `generated quantities` block produces posterior predictive samples for model checking and pointwise log-likelihoods for model comparison via LOO-CV.

Personality Moderation of Vocal Stress Responses

Research Question. The central question addressed by this model is whether the effect of exam-related stress on vocal acoustics (F0, NNE) is moderated by individual differences in personality pathology. Specifically, we ask: Do the five PID-5 domains—Negative Affectivity, Detachment, Antagonism, Disinhibition, and Psychoticism—differentially amplify or attenuate the stress-induced changes in vocal parameters during (a) anticipatory stress and (b) post-stressor recovery?

The Challenge: Personality Traits from Intensive Longitudinal Data. A key methodological challenge in this study concerns how personality traits are represented in the moderation analysis. Each participant completed approximately 20 EMA assessments over 2.5 months, providing repeated measures of each PID-5 domain. A naive approach would aggregate these observations into a single person-level mean for each domain and use these means as predictors in a standard multilevel regression. However, this approach discards valuable information and fails to account for measurement error:

the observed person-means are noisy estimates of the true latent traits, and treating them as known quantities underestimates uncertainty in the moderation effects.

Our model addresses this problem through a *joint measurement-and-outcome model* that simultaneously estimates latent personality traits from the EMA data and their moderating influence on vocal stress responses. This integrated approach has three key advantages:

1. **Measurement error correction**: Rather than using observed means as fixed predictors, the model estimates each participant's true latent trait score (θ_{id}) as a parameter, with appropriate uncertainty. This uncertainty propagates into the moderation estimates, yielding appropriately calibrated credible intervals.

2. **Borrowing strength across observations**: The repeated EMA measurements inform the latent trait estimates through a measurement model, allowing the model to distinguish stable trait variance from occasion-specific fluctuations.

3. **Coherent uncertainty quantification**: Because the latent traits and their effects are estimated jointly, the posterior distributions for moderation parameters (γ_1, γ_2) fully reflect uncertainty about both the traits themselves and their influence on vocal outcomes.

Model Specification. The model comprises two interconnected components: a *measurement model* for the EMA-based personality assessments and an *outcome model* for the vocal parameters.

Measurement Model (EMA). Let X_{nd} denote the observed score for participant i on domain d at EMA occasion n. The measurement model specifies:

$$X_{nd} \sim \text{Normal}(\theta_{i[n],d}, \sigma_d^{\text{ema}})$$

where θ_{id} is the latent true trait score for participant i on domain d, and σ_d^{ema} captures occasion-to-occasion variability (including both state fluctuations and measurement error). The latent traits are given standard normal priors:

$$\theta_{id} \sim \text{Normal}(0, 1)$$

This formulation treats each participant's approximately 20 EMA observations as repeated noisy indicators of a stable underlying trait, with the model learning both the trait estimates and the amount of occasion-level variability for each domain.

Outcome Model (Vocal Parameters). Let y_j denote the vocal outcome (F0 or NNE) for observation j, with participant index $i[j]$. Stress effects are parameterized using two orthogonal contrasts:

- c_1: Stress contrast (pre-exam vs. baseline);
- c_2: Recovery contrast (post-exam vs. pre-exam).

The outcome model specifies:

$$y_j \sim \text{Normal}(\mu_j, \sigma_y)$$

where the linear predictor μ_j includes fixed effects, random effects, and the crucial trait $\times$ contrast interactions:

$$\mu_j = \alpha + u_{i,1} + (\beta_1 + u_{i,2}) \cdot c_{1j} + (\beta_2 + u_{i,3}) \cdot c_{2j} + \sum_{d=1}^{5} \left[a_d \cdot \theta_{id} + \gamma_{1d} \cdot c_{1j} \cdot \theta_{id} + \gamma_{2d} \cdot c_{2j} \cdot \theta_{id} \right]$$

The parameters are:

- α: Grand mean (baseline vocal parameter);

- β_1, β_2: Population-average stress and recovery effects;
- $u_{i,1}$, $u_{i,2}$, $u_{i,3}$: Participant-specific random intercept and slopes;
- a_d: Main effect of trait d on baseline vocal level;
- γ_{1d}: **Stress moderation**—how trait d amplifies or attenuates the stress effect;
- γ_{2d}: **Recovery moderation**—how trait d shapes post-stressor trajectory.

The moderation parameters γ_{1d} and γ_{2d} are the quantities of primary theoretical interest. A positive γ_{1d} indicates that higher levels of trait d are associated with larger stress-induced changes in the vocal parameter.

Random Effects Structure. Participant-level random effects are specified using a non-centered parameterization for computational efficiency:

$$u_{i,k} = z_{i,k} \cdot \tau_k, \quad z_{i,k} \sim \text{Normal}(0, 1)$$

where τ_k are the random effect standard deviations. This structure allows for individual differences in baseline levels ($u_{i,1}$), stress reactivity ($u_{i,2}$), and recovery ($u_{i,3}$) beyond what is explained by the PID-5 traits.

Prior Specification. Priors were chosen to be weakly informative, incorporating domain knowledge about plausible parameter ranges while allowing the data to dominate inference (Table S9).

The priors on the moderation parameters (γ_{1d}, γ_{2d}) provide modest regularization, shrinking estimates toward zero in the absence of strong evidence. This helps guard against overfitting given the 10 moderation parameters (5 domains × 2 contrasts) being estimated.

Stan Implementation. The model was implemented in Stan. The complete code is available in the online repository accompanying this article.

Interpretation of Key Parameters. The model yields posterior distributions for 10 moderation parameters of primary interest:

Table S9

Prior specifications for the moderation model.

Parameter	Prior	Rationale
α	Normal(220, 30)	Centered on typical female F0 (Hz)
β_1, β_2	Normal(0, 10)	Allows stress effects up to $\pm$20 Hz
a_d	Normal(0, 5)	Modest trait effects on baseline
γ_{1d}, γ_{2d}	Normal(0, 3)	Regularization toward zero
τ_k	Exponential(0.5)	Weakly informative for SDs
σ_y	Exponential(0.1)	Allows residual SD up to $\sim$10 Hz
σ_d^{ema}	Exponential(1)	Weakly informative for EMA variability

- $\gamma_{1,\text{NegAff}}, \cdots, \gamma_{1,\text{Psych}}$: How each PID-5 domain moderates stress-induced vocal change (stress contrast $\times$ trait).
- $\gamma_{2,\text{NegAff}}, \cdots, \gamma_{2,\text{Psych}}$: How each domain moderates recovery-phase vocal change (recovery contrast $\times$ trait).

These parameters are expressed in the original units of the vocal outcome (Hz for F0, dB for NNE) per standard deviation of the latent trait. For example, $\gamma_{1,\text{NegAff}} = 3.0$ would indicate that a one-SD increase in latent Negative Affectivity is associated with an additional 3 Hz increase in F0 during the stress phase, beyond the population-average stress effect.

Posterior Predictive Checks. The `generated quantities` block produces posterior predictive samples (y_{rep}) for model checking. These were used to verify that the model adequately captured the distributional properties of the observed vocal data, including means, variances, and the pattern of individual differences across assessment phases.

MCMC Convergence Diagnostics

Estimation Procedure

All Bayesian models were estimated using Hamiltonian Monte Carlo (HMC) via Stan (Carpenter et al., 2017), accessed through the cmdstanr package in R. For the primary F0 moderation model, we ran 4 independent chains with 2,000 warmup iterations and 6,000 sampling iterations per chain, yielding 24,000 total post-warmup draws. Sampler adaptation parameters were set conservatively to ensure reliable exploration of the posterior: `adapt_delta = 0.99` and `max_treedepth = 15`. The same estimation procedure was applied to the NNE moderation model.

Convergence Assessment

Convergence was assessed using standard diagnostics recommended for HMC estimation (Vehtari et al., 2021):

1. **Divergent transitions**: No divergent transitions were observed in either model, indicating that the sampler explored the posterior without encountering regions of high curvature that would compromise inference.

2. **Maximum treedepth**: No iterations exceeded the maximum treedepth, suggesting that the sampler did not require truncation of its trajectory length.

3. **Energy Bayesian Fraction of Missing Information (E-BFMI)**: All chains showed E-BFMI values well above the recommended threshold of 0.2, indicating adequate exploration of the energy distribution.

4. $\hat{R}$ **(R-hat)**: The potential scale reduction factor was computed for all parameters. For the F0 model, all $\hat{R}$ values for key parameters (fixed effects, moderation coefficients, variance components) were below 1.01, and no parameters across the full model exceeded 1.05. Table S10 reports $\hat{R}$ values for key parameters.

5. **Effective Sample Size (ESS)**: Both bulk-ESS and tail-ESS were computed for all parameters. Bulk-ESS, which reflects the efficiency of sampling for posterior means and medians, exceeded 4,000 for all key parameters. Tail-ESS, which reflects sampling efficiency for posterior quantiles, exceeded 3,000 for all key parameters. These values substantially exceed the minimum recommended threshold of 400 (Vehtari et al., 2021).

Visual Diagnostics

Trace plots for all key parameters showed excellent mixing across chains, with no visible trends or differences between chains. Representative trace plots are shown in Figure S1.

Autocorrelation plots indicated low autocorrelation at lag 1 and negligible autocorrelation beyond lag 10 for all parameters, consistent with the high effective sample sizes reported above.

Posterior Predictive Checks

Model adequacy was assessed using posterior predictive checks (Gelman et al., 2013). Figure S2 shows the posterior predictive density overlaid on the observed F0 distribution. The model accurately captured the location, spread, and shape of the observed data.

Additional posterior predictive checks confirmed that the model adequately recovered:

- The mean F0 across all observations (observed: 192.5 Hz; 95% PPC interval: [191.2, 193.8])
- The standard deviation of F0 (observed: 22.1 Hz; 95% PPC interval: [20.8, 23.4])
- The pattern of means across assessment periods (Baseline $<$ Post-exam $<$ Pre-exam)

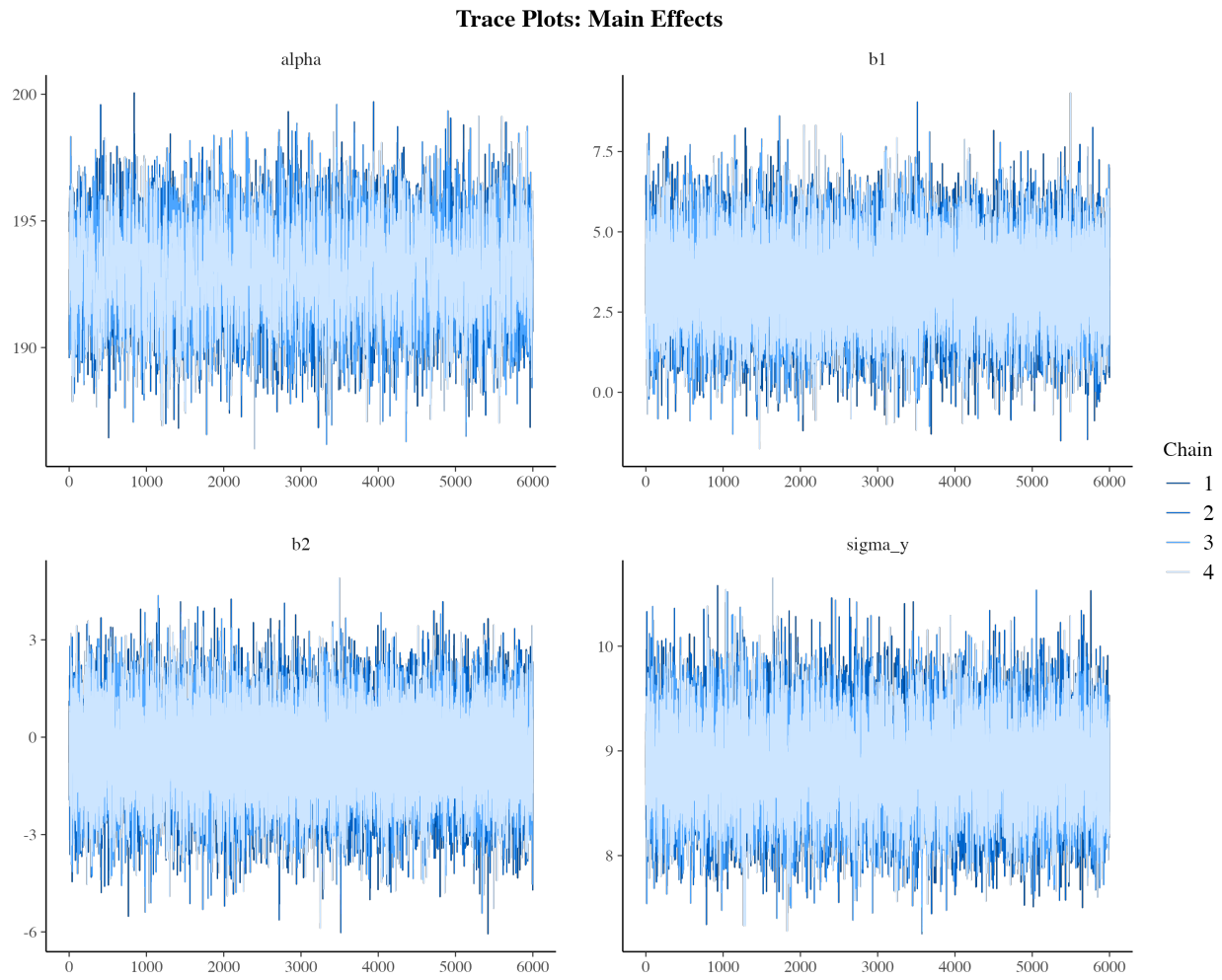

Figure S1. Trace plots for main fixed effects parameters showing good mixing across four chains.

Summary

All convergence diagnostics indicated that the MCMC sampler successfully explored the posterior distribution. The absence of divergent transitions, combined with $\hat{R}$ values below 1.01 and effective sample sizes exceeding 4,000, provides strong evidence that the posterior summaries reported in the main text are reliable. Posterior predictive checks confirmed that the model adequately captured the key features of the observed data.

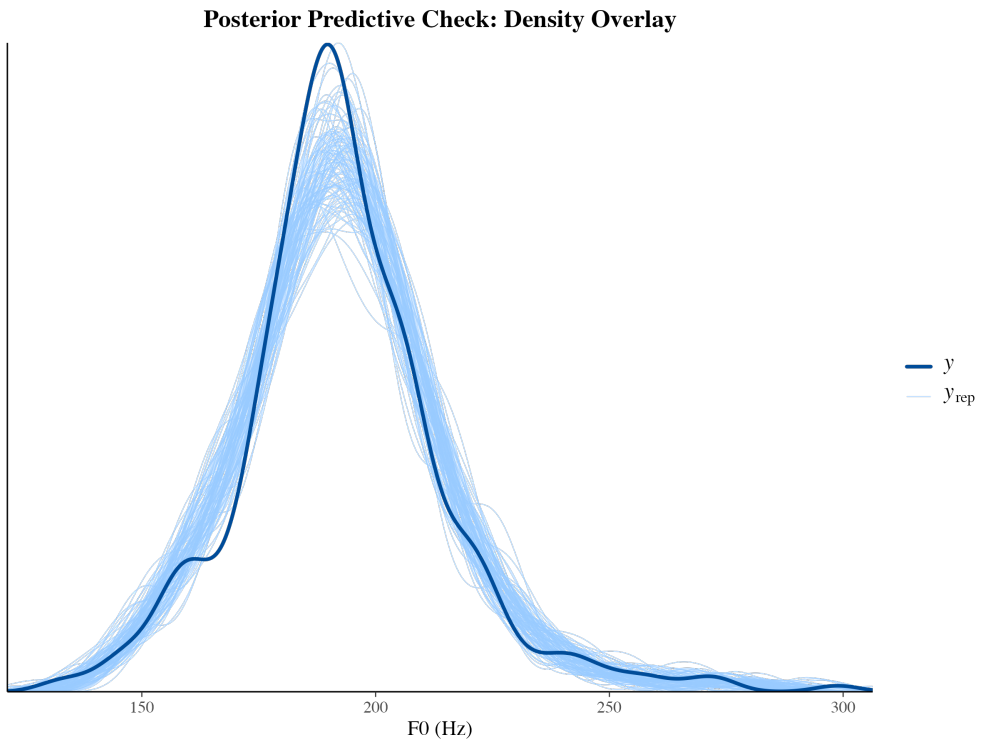

Figure S2. Posterior predictive check showing 100 draws from the posterior predictive distribution (light blue) overlaid on the observed F0 distribution (dark blue).

NNE Model Diagnostics

The same estimation and diagnostic procedures were applied to the NNE moderation model. Table S11 summarizes the convergence diagnostics for both models.

Table S11

Convergence diagnostics summary for F0 and NNE moderation models.

Model	Divergences	Max Treedepth Exceeded	$\hat{R}$ (max)	$\hat{R} > 1.01$	ESS Bulk (min)	ESS Tail (min)
F0	0	0	1.0015	0	4490	8441
NNE	0	0	1.0013	0	3775	4659

The NNE model showed equally good convergence properties: no divergent transitions, no iterations exceeding maximum treedepth, all $\hat{R}$ values below 1.01, and

minimum effective sample sizes well above recommended thresholds. Posterior predictive checks for the NNE model (Figure S3) confirmed adequate model fit.

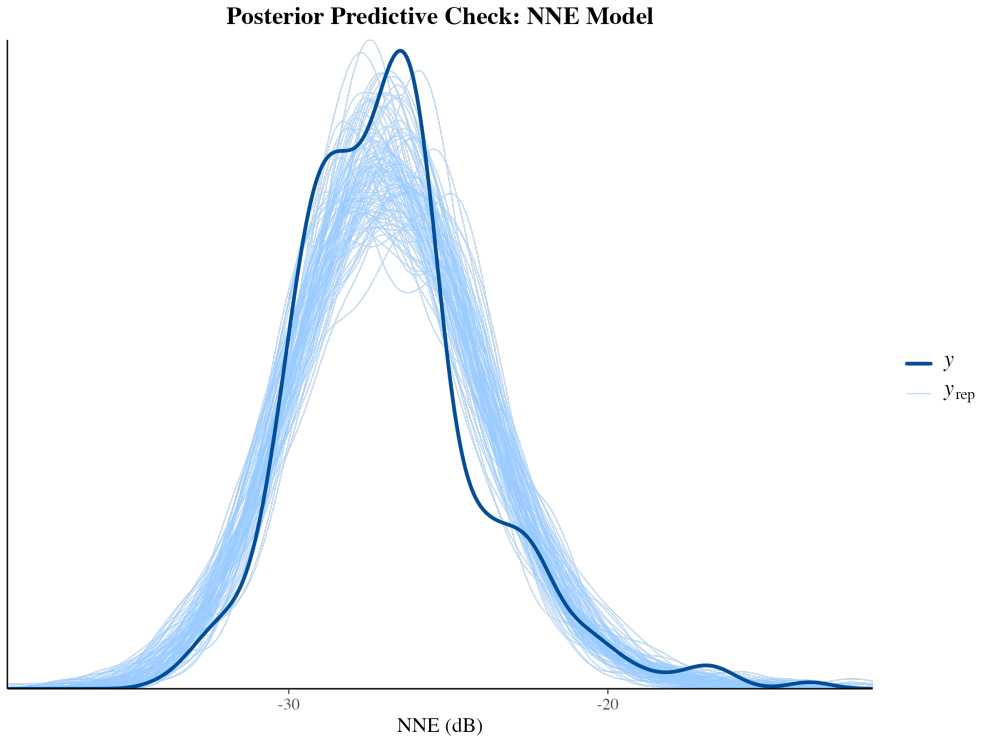

Figure S3. Posterior predictive check for the NNE moderation model.

Posterior Predictive Assurance of Moderation Effects

Rationale

Beyond estimating posterior distributions for the moderation parameters, we conducted a posterior predictive assurance analysis to quantify the probability that the *direction* of the observed moderation effect would replicate under the same study design. Rather than addressing statistical significance or interval exclusion criteria, this analysis focuses on directional replicability: given the fitted model and the uncertainty in its parameters, how likely is it that a new study with the same design would yield a moderation effect in the same direction?

This approach is conceptually distinct from frequentist power analysis. Instead of assuming a fixed but unknown true effect size, posterior predictive assurance integrates over the full posterior distribution of the model parameters, thereby reflecting uncertainty about both the magnitude of the effect and the data-generating process.

Procedure

Posterior predictive simulations were conducted using draws from the joint posterior distribution of the fitted moderation model. For each simulation, a new dataset was generated under the same design as the original study, including the same number of participants, the same stress and recovery contrasts, and a comparable distribution of EMA observations per participant. Latent personality traits for the new participants were drawn from their population prior distributions, consistent with the generative assumptions of the model.

For each simulated dataset, the moderation effect of interest—specifically, the interaction between Negative Affectivity and the stress contrast—was re-estimated using a fast proxy model. Replication success was defined using a minimal and directionally focused criterion: the estimated moderation effect was required to be positive (> 0). This

criterion captures whether the effect would replicate in direction, without imposing additional thresholds related to statistical significance or effect size magnitude. This process was repeated 1,000 times, yielding an empirical estimate of the posterior predictive probability of directional replication (assurance).

Results

Across posterior predictive replications, the moderation effect was positive in 92.1% of simulated datasets. The estimated posterior predictive probability of replication success was therefore 0.92, with a 95% credible interval of [0.90, 0.94], reflecting Monte Carlo uncertainty in the simulation-based estimate.

Interpretation

These results indicate a high probability that the direction of the moderation effect would replicate in a new sample drawn under the same design assumptions. In other words, given the fitted model and the uncertainty in its parameters, the interaction between Negative Affectivity and stress is expected to be positive in the large majority of replications.

At the same time, this analysis does not imply precise recovery of the effect magnitude. The posterior distribution of the moderation parameter remains relatively wide, indicating substantial uncertainty about the exact size of the effect. The assurance analysis therefore supports *directional robustness* of the moderation effect, while remaining agnostic about the degree of precision with which its magnitude can be estimated.

Taken together with the posterior estimates reported in the main analysis, these results suggest that the observed moderation effect is unlikely to be a chance reversal in direction, even though its quantitative strength should be interpreted with appropriate caution.

References

Bottesi, G., Caudek, C., Colpizzi, I., Iannattone, S., Palmieri, G., & Sica, C. (2024). Advancing understanding of the relation between criterion A of the alternative model for personality disorders and hierarchical taxonomy of psychopathology: Insights from an external validity analysis. *Personality Disorders: Theory, Research, and Treatment*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*, 1–32.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Lai, M. H. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods, 26*(1), 90–102.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718.

Table S10

Posterior summary and convergence diagnostics for key parameters (F0 moderation model).

Parameter	Mean	SD	$\hat{R}$	ESS (bulk)	ESS (tail)
Intercept (α)	192.95	1.78	1.001	4490	8441
Stress effect (β_1)	3.43	1.33	1.000	38288	19497
Recovery effect (β_2)	-0.49	1.34	1.000	38539	20141
γ_1 Negative Affectivity	3.14	1.68	1.000	39739	20045
γ_1 Detachment	-0.36	1.68	1.000	37929	20076
γ_1 Antagonism	-0.10	1.61	1.000	38621	18464
γ_1 Disinhibition	-0.21	1.84	1.000	37799	19097
γ_1 Psychoticism	-0.26	1.71	1.000	38639	19968
γ_2 Negative Affectivity	-0.45	1.70	1.000	37632	19706
γ_2 Detachment	-2.02	1.71	1.000	37665	19891
γ_2 Antagonism	3.16	1.61	1.000	37825	20302
γ_2 Disinhibition	-0.18	1.87	1.000	34125	19595
γ_2 Psychoticism	-1.42	1.70	1.000	39361	20089
Residual SD (σ_y)	8.85	0.44	1.000	14363	17189
Random intercept SD (τ_1)	18.81	1.27	1.000	5855	11332
Random stress slope SD (τ_2)	1.33	1.14	1.000	12845	13041
Random recovery slope SD (τ_3)	1.65	1.43	1.000	10546	13300