PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Emotional (in)stability: Neuroticism is associated with increased variability in negative emotion after all

Nina Mader[a,1], Ruben C. Arslan[a], Stefan C. Schmukle[a], and Julia M. Rohrer[a]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received July 19, 2022; accepted April 30, 2023

The personality trait neuroticism is tightly linked to mental health, and neurotic people experience stronger negative emotions in everyday life. But, do their negative emotions also show greater fluctuation? This commonsensical notion was recently questioned by [Kalokerinos *et al. Proc Natl Acad Sci USA* 112, 15838–15843 (2020)], who suggested that the associations found in previous studies were spurious. Less neurotic people often report very low levels of negative emotion, which is usually measured with bounded rating scales. Therefore, they often pick the lowest possible response option, which severely constrains the amount of emotional variability that can be observed in principle. Applying a multistep statistical procedure that is supposed to correct for this dependency, [Kalokerinos *et al. Proc Natl Acad Sci USA* 112, 15838–15843 (2020)] no longer found an association between neuroticism and emotional variability. However, like other common approaches for controlling for undesirable effects due to bounded scales, this method is opaque with respect to the assumed mechanism of data generation and might not result in a successful correction. We thus suggest an alternative approach that a) takes into account that emotional states outside of the scale bounds can occur and b) models associations between neuroticism and both the mean and variability of emotion in a single step with the help of Bayesian censored location-scale models. Simulations supported this model over alternative approaches. We analyzed 13 longitudinal datasets (2,518 individuals and 11,170 measurements in total) and found clear evidence that more neurotic people experience greater variability in negative emotion.

neuroticism | emotional variability | personality | censored regression

Significance

In everyday life, our emotions can change from moment to moment, and people experience such fluctuations to varying degrees. Psychologists have puzzled over the role that the personality trait neuroticism—a potent risk factor for mental illness—plays in such emotional variability. Do neurotic individuals experience not only stronger negative emotions but also more variability? This question resulted in controversy because it is methodologically challenging to separate effects of neuroticism on mean emotion from effects on variability. We suggested a different modeling approach to address the methodological issues, tested its performance on simulated data, and then reinvestigated a total of 13 longitudinal datasets. The findings suggest that more neurotic individuals indeed experience more variability in negative emotion in everyday life.

Emotions color our daily lives, and individuals reliably differ in their everyday emotional experiences. The personality trait neuroticism is associated with more frequent and stronger negative emotions both in everyday life and in people's responses to triggering circumstances (1). Neuroticism is therefore also referred to as negative emotionality (2). However, since the 1990s, studies have repeatedly suggested an additional link between neuroticism and emotion, specifically that higher neuroticism is associated with higher *variability* in emotion in daily life (3–5). As a consequence, the opposite pole of neuroticism is frequently labeled emotional stability.

But, the association between neuroticism and emotional variability has been contested. In 2020, Kalokerinos and colleagues (6) published an extensive meta-analysis of 11 datasets on the association between neuroticism and variability in negative emotion. Their meta-analysis suggested that previous studies may have detected a statistical artifact. The intensity of reported negative emotion in everyday life tends to be low in many samples, with the result that individuals often have average levels close to the lower bound of the response scale. But, if an individual's average negative emotion is close to the measurable minimum, there is very little room left for observing even lower values and, thus, little room for observing variability around the mean. The same would hold true for individuals whose average level of negative emotion is close to the upper bound of the scale, resulting in an inverse u-shaped dependency between the mean value and the potential variability around that mean (7). Thus, the observed associations between neuroticism and variability in negative emotion may be attributed to the well-established association between neuroticism and negative emotion: More neurotic individuals have higher mean levels, and thus, in the presence of bounded response scales, they simply have more "space" for variability around that mean.

Multiple approaches have been developed to address this dependency (8–11), but problems have been detected in all of them (6, 7). One prominent approach that was developed to solve these problems is the *relative variability index* (RVI) by Mestdagh and colleagues (7). To calculate the RVI, one first calculates the maximum possible variability one could have observed for a given mean value and response scale. The observed variability is then expressed as a percentage of the maximum possible variability, resulting in an index that lies between 0 and 1 for all possible mean values (except for mean values that lie on

Author affiliations: [a]Department of Psychology, University of Leipzig, Leipzig 04109, Germany

Author contributions: R.C.A., S.C.S., and J.M.R. designed research; N.M. performed research; N.M., R.C.A., and J.M.R. analyzed data; and N.M., R.C.A., S.C.S., and J.M.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: nina.mader@uni-leipzig.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2212154120/-/DCSupplemental.

Published May 30, 2023.

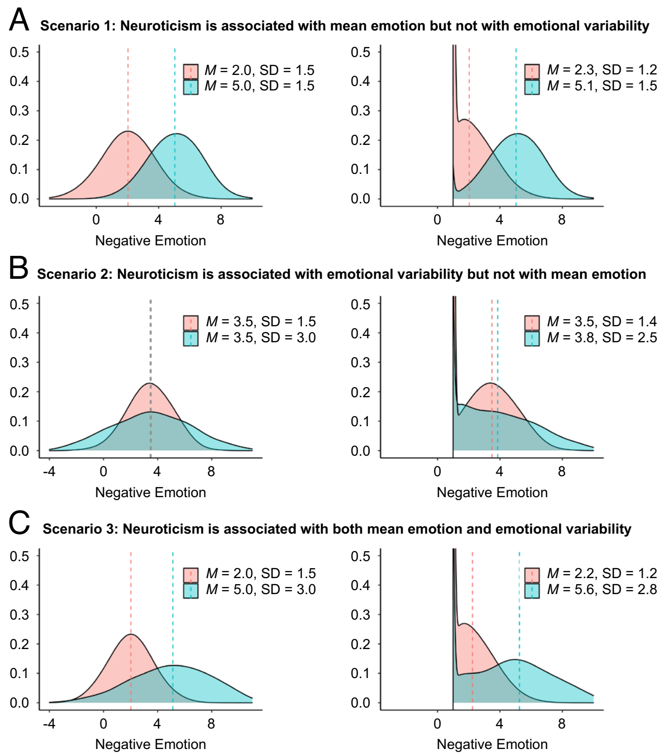

Fig. 1. Illustration of bounded scales in three different scenarios described in the text (*A*, *B*, *C*). Data from two individuals, one low and one high in neuroticism (red and green, respectively), each with 1,000 observations of negative emotion, were simulated. The true distributions are shown on the *Left*. The distributions on the *Right* show what is reproduced by a bounded scale when all values smaller than 1 (i.e., the lower limit of the scale) are counted as a value of 1. *M* = mean value, *SD* = standard deviation.

the scale boundary, where the maximum possible variability is zero). Kalokerinos et al. (6) meta-analyzed 11 datasets and found a very weak association between the RVI for negative emotion and the personality trait neuroticism, $r = 0.05$ (−0.01; 0.11). This value was not statistically different from zero. They thus concluded that neuroticism is not associated with more variability in negative emotion (6).

The RVI (7)—as well as other common attempts to statistically account for the dependency between the mean and variability—is justified by the assumption that differences in one's mean level of negative emotion lead to spurious differences in one's observed variability when scales are bounded. Such a scenario is depicted in Fig. 1*A*. We simulated the distributions of negative emotion across multiple measurements for two individuals: one low and one high in neuroticism. Here, we postulated that only the means but not the variability of the distributions of true negative emotion would vary between these two individuals (Fig. 1 *A*, *Left*). Due to the bounded scale, all values below the lower bound were truncated and added to the lower bound (Fig. 1 *A*, *Right*), resulting in heavy censoring and thus a more skewed distribution for the person low in neuroticism. In such a scenario, differences in mean values would indeed confound the association between neuroticism and variability, as they would introduce a spurious difference in the variability of observed emotion for these two individuals where there actually was none. Thus, adjusting the associations between neuroticism and variability for confounding by the central tendency (through the RVI or other methods) is sound in principle, although it is still possible to argue about how to best operationalize the variability (e.g., one suggestion was to count the number of episodes of negative emotion, ref. 12) and how to best operationalize the central tendency (e.g., one suggestion was to use the mode instead of the mean, ref. 13).

But, mean level and variability can also be correlated for other reasons. For example, consider two people with low and high levels of neuroticism who vary *only* in their emotional variability but not in their mean levels of negative emotion (Fig. 1 *B*, *Left*). Again, the values at the bottom end of the scale are truncated and added to the lower bound (Fig. 1 *B*, *Right*). In such a scenario, differences in emotional variability actually induce a spurious association between neuroticism and mean negative emotion; the confounding thus runs the other way. Here, the observed mean level of negative emotion is an *outcome* of differences in emotional variability—thus, statistical adjustment for it may even induce overcontrol bias (14) and move estimates away from the true association between neuroticism and emotional variability.

In reality, it may of course be the case that people with different neuroticism scores vary in both their underlying mean negative emotion and their emotional variability, and bounded scales distort both observed metrics (Fig. 1*C*). How then can we best use observed data—measured with limited scales—to make principled inferences about the true mean level of negative emotion and the true emotional variability as well as their associations with neuroticism?

Instead of statistical indices (e.g., the RVI) that are based on post hoc adjustments (which may work under restricted assumptions that are hard to deduce a priori), we suggest Bayesian censored location-scale models, which can model the (presumed) data-generating mechanism leading to excess zeroes. For our analyses, we used the *brms* package (15), which offers an easy way to fit these more complex models in R (16).

In our models, we treated observed negative emotion as a censored variable—thus, rather than trying to "work around" the scale bounds as is common in the existing literature, we explicitly incorporated them into our model. True emotion was assumed to follow an unbounded normal distribution that is censored during the measurement process. Thus, all observations that fell exactly at the lower bound of the scale were assumed to be indicative of values at *or below* the lower limit. This can account for heavily skewed distributions of *observed* emotion, which are often found in empirical data (17). On a substantive level, censoring may occur for multiple reasons. For example, the bounded response scale may make it impossible for respondents to accurately differentiate between low levels of individual negative emotions. But, censoring may also emerge on the aggregate level, as scales often focus on high-arousal negative emotions (18), thus potentially missing variability in more subtle negative emotions. Importantly, allowing for censoring does not impose additional assumptions, but rather removes the assumption that respondents who pick the lowest response option all experience the same level of negative emotion.

We also explicitly modeled censoring at the upper limit of the scale, but the upper limit had little impact on results for our research question because very few people ever report such high levels of negative emotion.

Furthermore, we used location-scale models in which all the parameters of the outcome distribution are allowed to vary, depending on the predictor variables. Thus, this framework allowed us to simultaneously estimate both the mean level of emotion and the variability around the said mean level as a function of neuroticism, taking into account uncertainty in *both* the mean and the variability.

To test the performance of our model in comparison with other approaches, we simulated data representing the three possible scenarios described previously (Fig. 1, association between neuroticism and the mean of emotion, association between neuroticism and the SD, and associations between neuroticism and both the mean and SD). Here, we simulated data for two distributions of emotion, one with strong left censoring (which corresponds to the observed distribution of negative emotion) and the other one with only little censoring (which more closely corresponds to the observed distribution of positive emotion). We then fit Bayesian censored location-scale models that allowed for effects of neuroticism on both the predicted mean and the predicted SD. We contrasted our model with multiple alternative approaches: Bayesian location-scale models without censoring (which were used in refs. 19–21; calculating the RVI (based on the SD) and correlating it with neuroticism (either without or with weights to account for error inflation (6, 7); and calculating the (regular) SD and correlating it with neuroticism (either without or with additional control for mean emotion; the latter two-step approach was prominent in older literature (10, 22, 23). As long as only a few observations were censored (i.e., for the distribution of positive emotion), all approaches converged on the same conclusion regarding the association between neuroticism and emotional variability. However, when many observations were censored (i.e., for the distribution of negative emotion), the RVI, the SD, and the Bayesian location scale model without censoring returned biased results, whereas Bayesian censored location-scale models returned the correct coefficients for all the three scenarios (Fig. 2).

In the scenarios in which neuroticism affected the mean but not the variability (Fig. 2*A*), as expected, the SD (with or without additional adjustment for the mean) and the Bayesian location-scale model without censoring detected spurious positive associations between neuroticism and variability. However, the RVI—which is supposed to correct for this flaw of the SD—actually shot past its target and detected a spurious negative association between neuroticism and variability, thus confirming the notion that this measure in some sense "overcorrects" (13). In the (admittedly unrealistic) scenario in which neuroticism affected the variability but not the mean (Fig. 2*B*), *all* approaches recovered the true association between neuroticism and variability. However, as expected, here a naïve analysis predicting mean emotion from

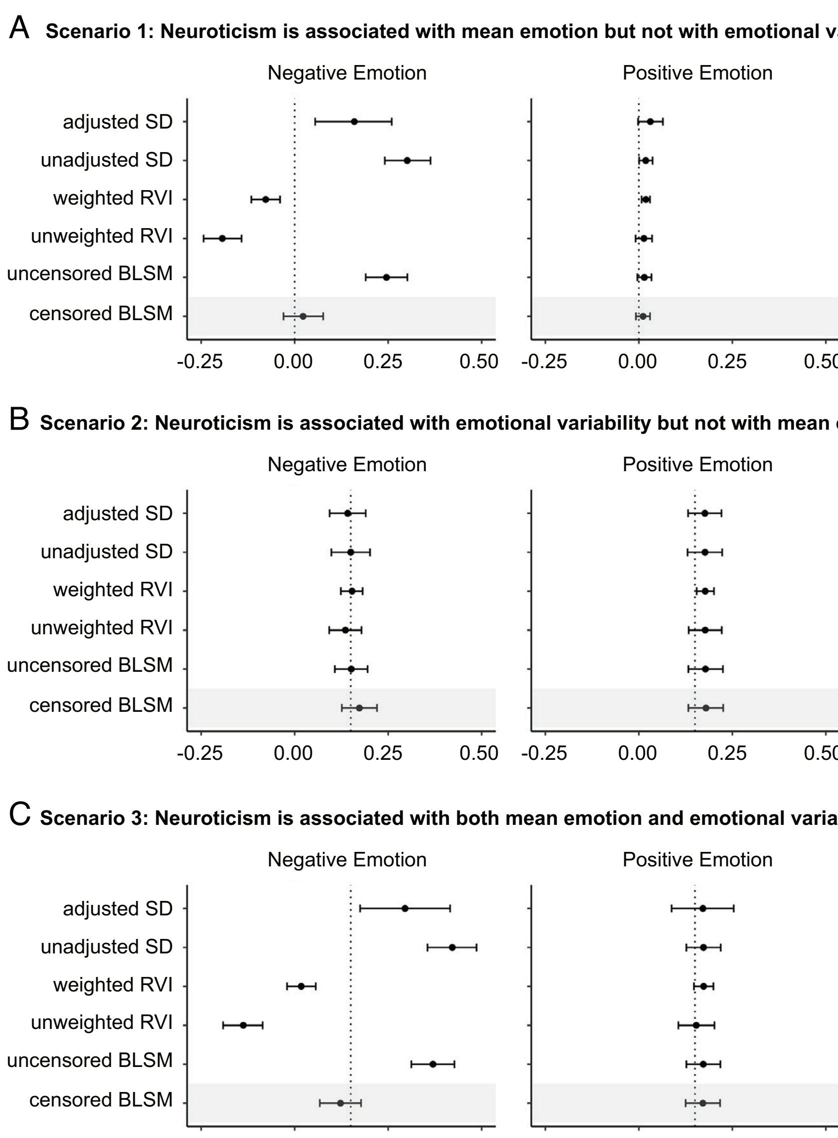

Fig. 2. Estimated associations between neuroticism and emotional variability in our three simulations (*A*, *B*, *C*) bounded by a 95% credible interval, depending on the statistical approach, separated by simulated negative (*Left*) and positive emotions (*Right*). BLSM = Bayesian location-scale model with and without censoring, RVI = Relative variability index unweighted and weighted, SD unadjusted and adjusted for the mean value, dotted line = true simulated parameter. Simulated data with N = 200 and daily levels of emotion over 30 d.

neuroticism (without taking into account potential differences in variability) resulted in a spurious positive association on mean emotion [with a coefficient of $b = 0.02$ (0.01; 0.03)]. By contrast, the Bayesian censored location-scale model correctly indicated no association between mean levels of emotion and neuroticism [$b = 0.00$ (−0.01; 0.01)]. Lastly, in the scenario in which neuroticism affected both the variability and the mean (Fig. 2*C*), the SD and the Bayesian location-scale model without censoring again consistently overestimated the association between neuroticism and emotional variability. The RVI consistently underestimated the association, and when applied without weighting, it actually even detected a spurious effect that *pointed in the opposite direction*. To see how the different approaches compare under an alternative data-generating mechanism, we also ran simulations in which emotion followed a skew normal distribution. These confirmed the overall pattern, including the suitability of our preferred model (*SI Appendix*, Part 8).

Overall, our simulations support Bayesian censored location-scale models as a data-analytic approach—but do these models actually change the conclusions about the (lack of an) association between emotional variability and neuroticism? We reanalyzed the 11 datasets included in the study by Kalokerinos et al. (6) and included two further large diary studies. In addition to Bayesian censored location-scale models, we also report results for the RVI, as it was developed with the aim of circumventing the dependence between the mean and variability of negative emotion, and to be able to compare our findings with those of Kalokerinos et al. (6). In a further association analysis, we weighted the RVI to account for residual error inflation at the scale end (7). Because both our simulations and the existing literature (6, 7, 24) have shown serious problems with the SD-based approaches, we did not investigate them further.

We meta-analyzed the association between neuroticism and negative emotional variability for each approach and additionally considered (variability in) positive emotion as a secondary outcome. Even though previous studies have focused on the association between neuroticism and variability in negative emotion, a lack of emotional stability could, in principle, imply higher variability in both negative *and* positive emotion. Lastly, researchers are often interested in variability as a predictor rather than as an outcome (19). Thus, we additionally ran analyses in which we used Bayesian censored location-scale models to extract estimates of negative emotional variability and subsequently tested whether these estimates predict neuroticism beyond mean negative emotion.

Results

Association between Neuroticism and Mean Negative Emotion. Unsurprisingly**,** our models confirmed a robust relationship between neuroticism and higher mean negative emotion scores in all the 13 studies, with coefficients ranging from $b = 0.13$ to $b = 0.41$, meta-analytic estimate: $b = 0.25$ [0.19; 0.30]. On average, for every 1-unit increase in the neuroticism score, the mean of negative emotion increased by 0.25 (Fig. 3*A*).

Association between Neuroticism and Negative Emotional Variability. Applying the RVI coefficient without weighting led to no associations between neuroticism and variability across all the 13 studies ($r = 0.02$ [−0.05; 0.08]; $b = 0.00$ [−0.09; 0.10]). These findings parallel the results reported by Kalokerinos and colleagues (6). Weighting the RVI to take into account residual error inflation at the scale end (7) led to narrower CIs but still close to inconclusive results [$b = 0.03$ (0.01; 0.04) and *SI Appendix*, Fig. S2].

By contrast, in Bayesian censored location-scale models, we found a consistent positive association between neuroticism and negative emotional variability (Fig. 3*B*). All studies showed positive coefficients, with an overall effect of $b = 0.10$ [0.07; 0.13] and negligible heterogeneity. To illustrate the magnitude of this estimate for a z-scored emotion variable (i.e., mean of 0, SD of 1), an individual with low (−1 SD) neuroticism would exhibit an SD of 0.9, whereas an individual with high (+1 SD) neuroticism would exhibit an SD of 1.1.

Gender as a potential confounder. We ran additional analyses controlling for gender. Because women tend to score higher on neuroticism, and they may also exhibit higher emotional variability, gender may offer an alternative explanation for the association between neuroticism and emotional variability. However, the association between neuroticism and negative emotional variability appeared only slightly diminished [$b = 0.09$ (0.06; 0.12) and *SI Appendix*, Fig. S3].

Positive emotion as a secondary outcome. For positive emotion, the results showed a clear relationship between higher levels of neuroticism and lower *mean* levels of positive emotion with a meta-analytic estimate of $b = -0.24$ (−0.29; −0.19). Considering positive emotional variability, the RVI without and with weighting as well as the Bayesian censored location-scale models all suggested quite small, heterogeneous effects. Meta-analytic estimates ranged from $b = 0.03$ to $b = 0.04$, and the 95% credible intervals mostly—but barely—included zero (*SI Appendix*, Fig. S4).

Negative emotional variability as a predictor of neuroticism. Flipping the logic of our analyses by predicting neuroticism from emotional variability (rather than estimating the effect of neuroticism on latent negative emotion and emotional variability), we found that negative emotional variability predicts neuroticism beyond mean negative emotion with a meta-analytic estimate of b = 0.32 (0.29; 0.35) (*SI Appendix*, Fig. S9).

Discussion

While it seems fairly intuitive that more neurotic individuals experience more variability in their emotions, methodological concerns have cast doubt on previous findings. In an impressive effort, Kalokerinos et al. (6) reanalyzed data from 11 studies using a statistical index RVI that is supposed to correct for any dependence between the mean of emotion and its variability. Here, we implemented an alternative modeling approach that, in a single step, allowed us to predict both the mean level and the variability of emotion while assuming a censored outcome variable. We thus explicitly took into account the possibility that emotional states outside of the scale range exist—they just cannot be reported faithfully when the lowest (or highest) response option has been reached.

Thanks to the groundwork laid by Kalokerinos et al. (6), we were able to apply our approach to a total of 13 studies, implementing diverse measures of emotion, and drawing on varied samples from different countries. Overall, we found clear evidence that neuroticism is associated with higher variability in negative emotion. Our estimates of the association could even be attenuated by measurement error, since the included measures of neuroticism were often fairly brief. This association aligns with our initial simulations, which showed that our model could recover an association that the RVI tends to underestimate or miss. Further in line with our simulations, for positive emotion, the results converged across approaches—highlighting how the performance of methods for statistical adjustment crucially depends on the censored distribution of the variable of interest. Findings suggested at best a small association between neuroticism and variability in positive emotion,

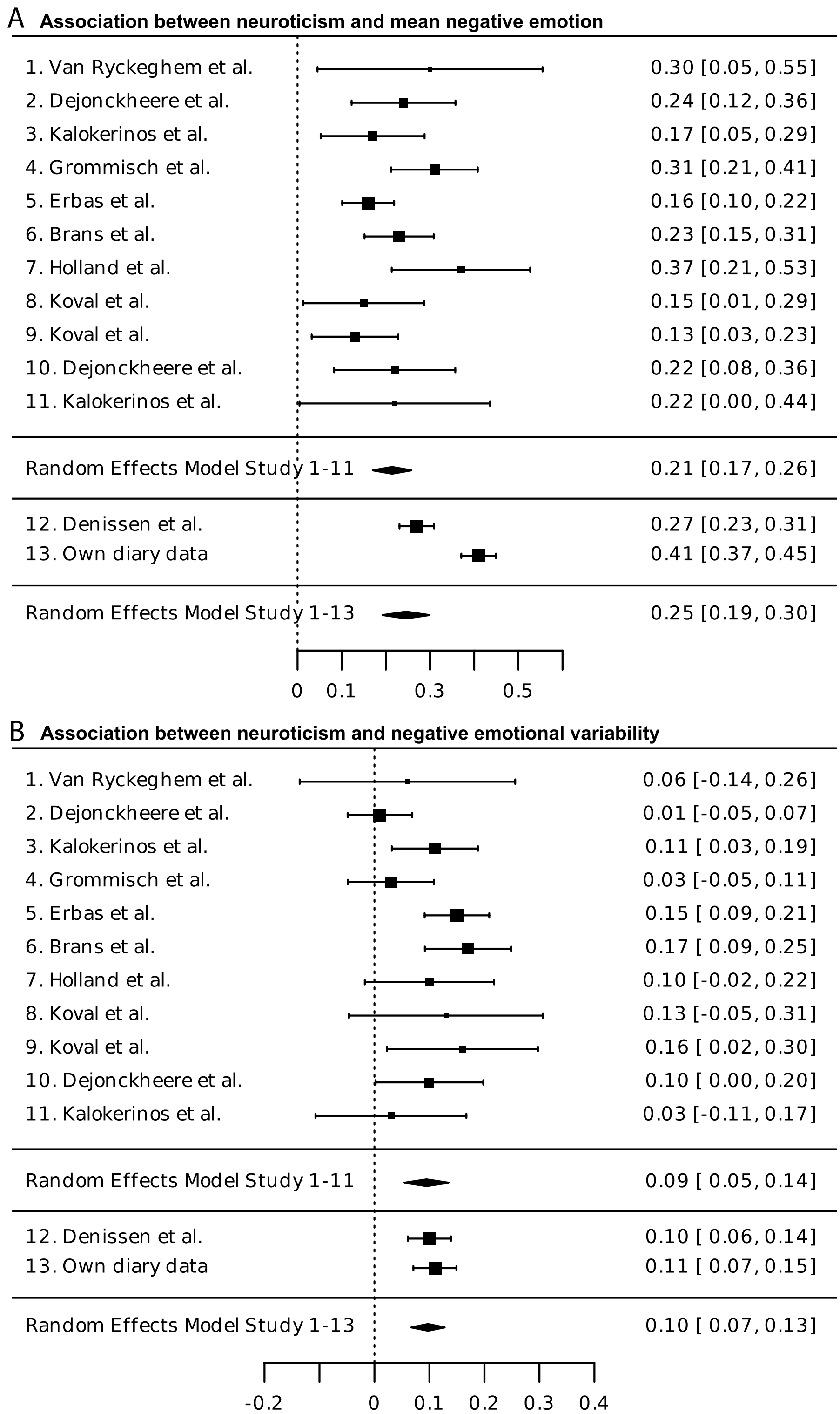

Fig. 3. Meta-analysis of the association (slope coefficients) between neuroticism and both mean levels of negative emotion (*A*) and negative emotional variability (*B*) using Bayesian censored location-scale models. Studies 1 to 11 are those analyzed by Kalokerinos et al. (6). Error bars show the 95% credible intervals.

which may potentially be heterogeneous across studies. Taken together, our results show that taking modeling seriously is not merely methodological nitpicking but can make a difference for substantive conclusions—such as the question of whether or not neuroticism is indeed associated with lower emotional stability.

Our modeling approach was quite light handed and still relies on the ubiquitous assumption of normally distributed variables; we merely considered that emotions are censored when using bounded response scales, which can conceal plausible variability, and which is sufficient to reverse substantive conclusions. While censored location-scale regression models are quite versatile, they are not meant to be a universal solution to all research questions involving variability—the choice of model should be motivated by the research question at hand (e.g., if variability is meant to be a predictor rather than just an outcome) as well as by our understanding of the dynamics of emotions.

Future efforts to improve our understanding of the dynamics of emotions should aim to model the data-generating processes underlying self-reports of emotion more comprehensively and under different sets of assumptions. For example, when instructed to recall emotions over a certain period of time (since the last experience sampling occasion, over the last day, over the last couple of days), do respondents plausibly aggregate over the time period of interest, or may there already be interindividual differences in recall processes associated with neuroticism? Furthermore, we focused on aggregated emotion scales, but moving to the item level may provide further insights closer to actual response behavior and help elucidate the question whether respondents can reliably report the absence of individual emotions. Thanks to the flexibility of Bayesian modeling, this could be implemented within the same framework as applied here (25). Descriptive work on time series suggests that individuals' emotions may follow a multimodal distribution [i.e., a distribution with multiple peaks, (17)]. Such a distribution may call for a more complex data-generating model—or alternatively suggests that some types of response formats (e.g., sliders with a default value) exhibit poor psychometric properties. Lastly, moving from recognition-based to recall-based measures of emotion may result in quite different models that provide more insights about the occurrence of negative emotion in everyday life (18).

Taking modeling more seriously thus highlights gaps in our understanding of how self-reports come into existence; how people choose which rating to give. It pushes us toward better-specified theories (26) and the additional investigation of cognitive processes (27). It may eventually even result in improved measurement approaches to provide a fuller picture of emotional variability and its correlates.

Materials and Methods

Simulated Data. To test the performance of the Bayesian censored location-scale model in comparison with the variability indices, we began with simulated data, ensuring that we knew the ground truth and could thus determine which analysis recovered the parameters of interest. For each of the three scenarios described in the main text [association between neuroticism and (a) the mean of emotion, (b) the SD of emotion, (c) both; see Fig. 1], we simulated a dataset consisting of 200 individuals with 30 observations each. Each person i was assigned a neuroticism score, a person-specific random intercept for mean emotion ($b_{0\_M\_i}$), and a person-specific random intercept for the SD around their mean ($b_{0\_SD\_i}$), all of which were drawn from standard normal distributions. Emotion was simulated for each scenario following the formula

$$X \sim N[\mu i = b0\_M + b0\_M\_i + b1\_M * \text{neuroticism}_i,$$

$$\sigma i = \exp(b0\_SD + b0\_SD\_i + b1\_SD * \text{neuroticism}_i)],$$

with global intercepts ($b_{0\_M,}$ b_{0_SD}). The exponential function ensured that the SD could take on only positive values.

We then censored the simulated emotion variable by replacing every value below the lower bound of the scale (1) and above the upper bound of the scale (5) with the respective scale bound.

We used the same slope parameters for simulations of positive and negative emotion:

- Scenario 1: Neuroticism is associated with mean levels of emotion but not with emotional variability, $b_{1\_M} = 0.5$ and $b_{1\_SD} = 0$
- Scenario 2: Neuroticism is associated with emotional variability but not with mean levels of emotion, $b_{1\_M} = 0$ and $b_{1\_SD} = 0.15$
- Scenario 3: Neuroticism is associated with both mean levels of emotion and emotional variability, $b_{1\_M} = 0.5$ and $b_{1\_SD} = 0.15$.

The only difference between the data-generating models for positive and negative emotion was the global intercept of the mean; we shifted the whole distribution to the left for negative emotion, resulting in more censoring at the lower end. Parameters were chosen so that the simulated distributions resembled typically observed distributions of negative and positive emotion. We then analyzed the simulated data following the analysis procedure outlined below. For detailed simulation results, see Fig. 2. The code for the simulation can be found at https://osf.io/9dcxn/.

Empirical Data. Datasets S1–S11 were previously reanalyzed and made available by Kalokerinos et al. (6) (https://osf.io/gvfdx/). In addition, we used Denissen and Kühnel's (28) data, which were also freely available (https://osf.io/6ghcx/), and a dataset we collected (29) (https://osf.io/k5zmd/). All data used were already deidentified prior to our analysis. Details about each study are presented in *SI Appendix*, Table S1.

Emotion. The included studies used different measures of negative and positive emotion. All but one sample (Study 11) used multiple items per dimension (details are presented in *SI Appendix*, Table S2). As different response scales were used, we decided to rescale all given answers to a scale ranging from one to five to facilitate the interpretation of coefficients. Descriptive statistics can be found in *SI Appendix*, Fig. S1 and Table S4 displays the observed distributions of emotion.

Neuroticism.

Big five inventory (BFI). Seven studies used the BFI (2) to assess personality, including neuroticism. Different versions of the BFI were used in different languages. More details about the scales and descriptive statistics are shown in *SI Appendix*, Tables S3 and S4.

Ten-item personality inventory (TIPI). Six studies used by Kalokerinos et al. (6) used the TIPI; ref. 30, which assesses neuroticism with two items, rated on a scale ranging from one to seven: "I see myself as anxious, easily upset" and "I see myself as calm, emotionally stable" (reverse coded), which were averaged. All neuroticism measures were rescaled to a common response scale of one to five.

Analyses. Analyses were performed with an identical analysis pipeline for each study, for both negative and positive emotions. All details can be found in the code, which is available here: https://osf.io/9dcxn/. For each dataset and each measure of emotion (negative emotion, positive emotion), we ran a *Bayesian censored location-scale model* and additionally estimated the association between neuroticism and emotional variability with the unweighted and the weighted RVI.

Bayesian censored location-scale regression models. We used *brms* (15), which implements Bayesian multilevel models using an R interface to the probabilistic programming language *Stan* (31). The Bayesian censored location-scale model allows all response distribution parameters to be predicted at the same time. While the mean (*location*) is modeled with the help of the identity link function (which is also the case in regular linear regression models), the residual SD (*scale*) is modeled with a log-link to ensure that only positive values are predicted.

Observed emotion, the dependent variable, was modeled as (left or right) censored when the observed values were at the (lower or upper) bound of the scale. Neuroticism was a predictor of both the mean and the SD of emotion. To account for the nested structure of the data, we also included random person intercepts on both the mean and the SD.

From the estimated models, we extracted the slope coefficients of interest (association between neuroticism and the mean, association between neuroticism and the SD) and conducted random-effects meta-analyses across studies. Meta-analytic estimates were calculated first for all studies that were part of the study by Kalokerinos et al. (6) and second for all of the 13 studies that we had available.

Model evaluations. We used graphical posterior predictive checking to examine the model's fit more closely. This involves comparing the data generated by the model to the observed data (*SI Appendix*, Fig. S5).

Additionally, we evaluated different assumptions about the residual SDs. First, a much simpler model which assumed a constant residual term (i.e., a regular multilevel regression model) always fit worse than our main model which allowed for heterogeneity of the residual SD, highlighting the need to model emotional variability. Second, a more complex model which additionally allowed heterogeneity in mean emotion to vary with neuroticism only led to small improvements in model fit for negative emotions and did not change the focal estimate of the relationship between neuroticism and individuals' emotional variability (*SI Appendix*, Part 5).

Furthermore, to gauge the extent to which modeling the outcome as censored matters, we reran the analyses of the Bayesian Location-Scale Model, but without censoring (*SI Appendix*, Part 6). As in our simulations, effects on the residual SD tended to be exaggerated when censoring was ignored [$b = 0.17$ (0.12; 0.21) compared to $b = 0.10$ (0.07; 0.13) when allowing for censoring, see *SI Appendix*, Part 6 for details].

SD. For the simulated data, we calculated the within-person SD for each person for positive and negative emotions. In a second step, the SD was used as the dependent variable in a regression analysis in which simulated neuroticism was the sole predictor. We log-transformed the SD to make the estimates comparable to the censored location-scale models, which used a log link. In a third step, we additionally adjusted the regression for the observed mean of emotion.

RVI. To calculate the RVI, we used the *relative Variability* R package (Version 1.0, ref. 32). The RVI can be applied to several variability measures (7). Like Kalokerinos et al. (6), we used SD to calculate the RVI as the proportion of the observed SD relative to the maximum SD that would be possible given a certain mean ($\frac{SD_i}{\max(SD_i \mid M_i)}$). Participants with mean values that coincided with the bounds of the scale were excluded because, in this case, the maximum possible variability was zero and division by zero is not allowed (see *SI Appendix*, Table S5, for the final sample sizes used in the analyses with the RVI). As the RVI shows an unintended amplification of errors for values near the scale bounds, the developers of the index suggested a weighting method by which more precise values are assigned more weight (7). Weighting is done using the inverse of the inflation factor (see *SI Appendix,* Fig. S2 for a brief illustration and discussion of how weighting affects results).

For the RVI, we ran simple correlations and both unweighted and weighted regression analyses. In the regression analyses, we log-transformed the RVI to make the regression estimates comparable to the estimates of the Bayesian censored location-scale models, which use a log link. We then conducted a random-effects meta-analysis across all the 13 studies.

Negative emotional variability as a predictor of neuroticism. For these additional analyses, we ran Bayesian censored location-scale regression models as described above, but omitted neuroticism as a predictor. We used these models to extract estimates of mean negative emotion and negative emotional variability for each individual and subsequently used both estimates to predict neuroticism. We then once again conducted random-effects meta-analysis across all studies to assess whether emotional variability uniquely predicts neuroticism, controlling for mean emotion (see *SI Appendix*, Part 7 for more details). In a straightforward extension of our main analyses, this approach allows us to test to which extent model-estimated emotional variability predicts neuroticism, over and above mean emotion. However, this two-step approach does not optimally propagate uncertainty, and researchers who are interested in research questions involving emotional variability as a predictor should instead implement a structural equation modeling approach (e.g., ref. 21) in software such as Stan, treating emotion as a censored latent variable.

Data, Materials, and Software Availability. Anonymized Data and Code data have been deposited in Open Science Framework https://osf.io/k5zmd/ (29).

1. B. B. Lahey, Public health significance of neuroticism. *Am. Psychol.* **64**, 241–256 (2009).
2. O. P. John, L. P. Naumann, C. J. Soto, "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues" in *Handbook of Personality: Theory and Research*, O. P. John, R. W. Robins, L. A. Pervin, Eds. (The Guilford Press, ed. 3rd, 2008), pp. 114–158.
3. L. Hepburn, M. W. Eysenck, Personality, average mood and mood variability. *Pers. Individ. Diff.* **10**, 975–983 (1989).
4. G. Murray, N. B. Allen, J. Trinder, Longitudinal investigation of mood variability and the ffm: Neuroticism predicts variability in extended states of positive and negative affect. *Pers. Individ. Diff.* **33**, 1217–1228 (2002).
5. B. Dauvier, J.-B. Pavani, S. Le Vigouroux, J.-L. Kop, A. Congard, The interactive effect of neuroticism and extraversion on the daily variability of affective states. *J. Res. Pers.* **78**, 1–15 (2019).
6. E. K. Kalokerinos *et al.*, Neuroticism may not reflect emotional variability. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9270–9276 (2020).
7. M. Mestdagh *et al.*, Sidelining the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychol. Methods* **23**, 690–707 (2018).
8. B. M. Baird, K. Le, R. E. Lucas, On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *J. Pers. Soc. Psychol.* **90**, 512–527 (2006).
9. M. Eid, E. Diener, Intraindividual variability in affect: Reliability, validity, and personality correlates. *J. Pers. Soc. Psychol.* **76**, 662–676 (1999).
10. P. Koval, M. L. Pe, K. Meers, P. Kuppens, Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion* **13**, 1132–1141 (2013).
11. J. Aitchison, S. M. Shen, Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261–272 (1980).
12. M. Wenzel *et al.*, Variability in negative affect is an important feature of neuroticism above mean negative affect once measurement issues are accounted for. *Eur. J. Pers.*, 1–14 (2022).
13. W. R. Ringwald, A. G. C. Wright, "Overcoming the confound of means and variability for measuring everyday emotion dynamics related to neuroticism". PsyArXiv [Preprint] (2022). 10.31234/osf.io/nxbyd (Accessed 1 June 2022).
14. J. M. Rohrer, Thinking clearly about correlations and causation: Graphical causal models for observational data. *Adv. Methods Pract. Psychol. Sci.* **1**, 27–42 (2018).
15. P.-C. Bürkner, **brms**: An *R* package for Bayesian multilevel models using *Stan*. *J. Stat. Soft.* **80**, 1–28 (2017).
16. R Core Team, *R: A language and environment for statistical computing 4.1.1* (R Foundation for Statistical Computing, Vienna, Austria, 2022).
17. J. M. B. Haslbeck, O. Ryan, F. Dablander, "Multimodality and skewness in emotion time series". PsyArXiv [Preprint] (2022). 10.31234/osf.io/qudr6 (Accessed 15 June 2022).
18. Y. Li, A. Masitah, T. T. Hills, The emotional recall task: Juxtaposing recall and recognition-based affect scales. *J. Exp. Psychol. Learn. Mem. Cog.* **46**, 1782–1794 (2020).
19. G. C. Hisler, Z. Krizan, T. DeHart, A. G. C. Wright, Neuroticism as the intensity, reactivity, and variability in day-to-day affect. *J. Res. Pers.* **87**, 103964 (2020).
20. D. R. Williams, S. R. Martin, S. Liu, P. Rast, Bayesian multivariate mixed-effects location scale modeling of longitudinal relations among affective traits, states, and physical activity. *Eur. J. Psychol. Assess.* **36**, 981–997 (2020).
21. P. Rast, E. Ferrer, A mixed-effects location scale model for dyadic interactions. *Multivariate Behav. Res.* **53**, 756–775 (2018).
22. U. W. Ebner-Priemer, M. Eid, N. Kleindienst, S. Stabenow, T. J. Trull, Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *J. Abnorm. Psychol.* **118**, 195–202 (2009).
23. M. Eid, E. Diener, Intraindividual variability in affect: Reliability, validity, and personality correlates. *J. Pers. Soc. Psychol.* **76**, 662–676 (1999).
24. W. Kalmijn, R. Veenhoven, Measuring inequality of happiness in nations. In search for proper statistics. *J. Happiness Stud.* **6**, 357–396 (2005).
25. E. Donnellan, S. Usami, K. Murayama, "Random item slope regression: Examining both similarities and differences in the association with individual items". PsyArXiv [Preprint] (2022). 10.31234/osf.io/s6erz (Accessed 20 June 2022).
26. R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Chapman and Hall/CRC, ed. 2, 2020).
27. A. M. Wood, G. D. A. Brown, J. Maltby, P. Watkinson, How are personality judgments made? A cognitive model of reference group effects, personality scale responses, and behavioral reactions: Relative personality judgments. *J. Pers.* **80**, 1275–1311 (2012).
28. J. J. A. Denissen, A. Kühnel, Berlin diary study at Humboldt Universität zu Berlin (2008). Data retrieved from. https://osf.io/6ghcx/. Accessed 20 November 2021.
29. J. M. Rohrer, I. S. Seifert, R. C. Arslan, S. C. Schmukle, Happiness Diary Data: Affect Variability (2018). Data retrieved from https://osf.io/k5zmd/. Accessed 10 March 2021.
30. S. D. Gosling, P. J. Rentfrow, W. B. Swann, A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).
31. Stan Development Team, Stan user's guide 2.31 (2022). https://mc-stan.org. Accessed 15 November 2021.
32. M. Mestdagh, Relative variability R package version 1.0 (2016). Retrieved from. https://ppw.kuleuven.be/okp/software/relative_variability/Link/. Accessed 25 March 2021.