AMERICAN PSYCHOLOGICAL ASSOCIATION

Emotion


ISSN: 1528-3542

2023, Vol. 23, No. 8, 2117–2141
https://doi.org/10.1037/emo0001218

# Multimodality and Skewness in Emotion Time Series

Jonas Haslbeck[1, 2], Oisín Ryan[3], and Fabian Dablander[2]
[1] Department of Clinical Psychological Science, Maastricht University
[2] Department of Psychological Methods, University of Amsterdam
[3] Department of Methodology and Statistics, Utrecht University

The ability to measure emotional states in daily life using mobile devices has led to a surge of exciting new research on the temporal evolution of emotions. However, much of the potential of these data still remains untapped. In this paper, we reanalyze emotion measurements from seven openly available experience sampling methodology studies with a total of 835 individuals to systematically investigate the modality (unimodal, bimodal, and more than two modes) and skewness of within-person emotion measurements. We show that both multimodality and skewness are highly prevalent. In addition, we quantify the heterogeneity across items, individuals, and measurement designs. Our analysis reveals that multimodality is more likely in studies using an analog slider scale than in studies using a Likert scale; negatively valenced items are consistently more skewed than positive valenced items; and longer time series show a higher degree of modality in positive and a higher skew in negative items. We end by discussing the implications of our results for theorizing, measurement, and time series modeling.

*Keywords:* emotion time series, multimodality, skewness, response scales, experience sampling methodology



Emotions allow us to learn from our environment and respond to it, and therefore evolve dynamically across time (Frijda, 2017; Kuppens & Verduyn, 2017). Consequently, a full understanding of emotion dynamics requires data at the timescale at which emotions evolve (Davidson, 1998; Schimmack et al., 2000; Verduyn et al., 2011). There has been a long-standing interest in studying within-person emotion dynamics (e.g., Lebo & Nesselroade, 1978), however, the possibility to measure peoples' emotions in daily life with mobile devices has led to a surge of new research (e.g., Kuppens et al., 2022), addressing fundamental questions such as whether emotions are discrete or continuous (Zelenski & Larsen, 2000), how often emotions occur and how they co-occur (Trampe et al., 2015), and how emotion dynamics relate to psychopathology (Kuppens et al., 2010; Wichers et al., 2015).

While time series data have already been leveraged to gain important new insights into emotion dynamics, many phenomena critical to studying emotions still remain unexplored. So far, most studies have focused on means, (co)variances, and relationships across time points (e.g., Bringmann et al., 2013; Trampe et al., 2015; Zelenski & Larsen, 2000), but little attention has been paid to the distributional form of emotion measurements. Specifically, the number of modes (i.e., "peaks") and the skewness of emotions in daily life are characteristics that are highly relevant for theory, measurement, and data analysis. In this paper, we for the first time systematically assess modality and skewness in emotion time series.

The modality and skewness of emotion measurements critically inform emotion dynamics. For example, a distribution with two modes would imply that an individual experiences an emotion in two "states" with some variations around those states: a person might experience almost no anger most of the time, but becomes angry in some situations, in which the person switches to the alternative state with a heightened intensity of anger. In contrast, a symmetric unimodal distribution would imply that the person experiences a typical intensity of anger all of the time, and varies around this typical intensity based on events in their environment. In addition, interindividual differences in modality and skewness may be related to characteristics such as personality, emotion regulation strategies, or psychopathology. Through such associations on the between-person level, analyzing modality and skewness might provide new insights into the heterogeneity of emotion and emotion regulation and the etiology of mental disorders (Lincoln et al., 2022).

This article was published Online First May 11, 2023.

We would like to thank the authors whose data we reanalyzed for making their data openly available. Without their commitment to open science, this paper would not have been possible. In addition, we would like to thank Denny Borsboom, Laura Bringmann, Julian Burger, Aaron Fisher, Eiko Fried, Jens Lange, Han van der Maas, Disa Sauter, Leonie Vogelsmeier, Charlotte Vrijen, Lourens Waldorp, and Leon Wendt for helpful discussions and comments on earlier versions of this paper. Jonas Haslbeck was supported by NWO Vici Grant No. 181.029 and by the project "New Science of Mental Disorders" (https://www.nsmd.eu/), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation Grant 024.004.016). Oisín Ryan was supported by ERC Consolidator (Grant 865468).

All authors contributed equally to this paper and should be considered joint first authors.

All data and code to reproduce our results are available from http://github.com/jmbh/EmotionTimeSeries and http://github.com/jmbh/ModalitySkewnessPaper.

Jonas Haslbeck, Oisín Ryan and Fabian Dablander contributed equally to conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing–original draft, writing–review and editing, and data curation.

Correspondence concerning this article should be addressed to Jonas Haslbeck, Department of Psychological Methods, University of Amsterdam, Psychological Methods, Nieuwe Achtergracht 129-B, Postbus 15906, 1018 WT, Amsterdam, the Netherlands. Email: jonashaslbeck@protonmail.com

Since modality and skewness of emotions are highly relevant theoretically, we need to ensure that our study designs allow us to appropriately capture them. Currently, a large variety of measurement tools exist in studies using the experience sampling methodology (ESM; Brose et al., 2020; Trull & Ebner-Priemer, 2020). Responses are measured on a Likert scale (e.g., with five or seven categories) or on a continuous visual analogue slider; response categories are labeled in different ways; in some studies, ticks on sliders are visualized; and sometimes a slider is initialized at a certain initial value. It is therefore important to determine which measurement design choices are best suited to capture modality and skewness, but also to ensure that the measurement tools themselves do not induce any particular distributional form.

Finally, the distributional form of data has major ramifications for statistical modeling. The currently most popular statistical model for emotion time series—the vector autoregressive (VAR) model (e.g., Bringmann et al., 2013; Hamaker et al., 2015; Vanhasbroeck et al., 2021)—only fits symmetric unimodal distributions (e.g., Hamilton, 1994). If emotion measurements turn out to deviate considerably from unimodal distributions, this would mean that such models would be gravely misspecified. This implies that they provide a poor fit to the data, but also that interpreting their parameters may be misleading (Haslbeck & Ryan, 2022). The presence of, for example, multimodality would suggest the need for other time series models that capture such behavior, such as regime-switching or hidden Markov models (HMMs; e.g., Visser & Speekenbrink, 2010).

Here, we present the first systematic investigation of modality and skewness in emotion time series measurements. We use seven public emotion ESM datasets (Bringmann et al., 2013, 2016; Fisher et al., 2017; Fried et al., 2020; Rowland & Wenzel, 2020; Vrijen et al., 2018; Wright et al., 2017), containing in total 11,520 univariate time series of 55 unique emotion items from 835 individuals, captured by a variety of different study designs. We show that both multimodality and skewness are highly prevalent across most of these studies. In fact, using a conservative classification method, less than half of emotion measurements (14%–40% across studies) exhibited a unimodal symmetric distribution. We show that the modality and skewness of emotion measurements are highly heterogeneous across items, individuals, and measurement designs. Our analysis reveals that multimodality is much more likely to be observed in studies using an visual analogue slider scale compared to a discrete Likert scale; that negatively valenced items are consistently more skewed than positively valenced items; and that longer time series tend to show higher modality in positive items, and higher skewness for negative items. We end by discussing the implications of our results for theorizing, measurement, and data analysis in the field of emotion dynamics along with practical recommendations for applied researchers.

## Modality and Skewness in Seven Emotion Time Series Datasets

We assess the distributional form of empirical emotion measurements in seven openly available ESM datasets, which we introduce in the "Data" section. In the "Assessing Modality and Skewness" section, we introduce the heuristic method we use to determine the qualitative form of the distribution in terms of the number of modes present and, for unimodal densities, as skewed or symmetric. In the "Results" section, we report the distributional form of emotion items across studies, quantify heterogeneity on the levels of items, individuals, and studies, and explain some of this heterogeneity with item characteristics, person characteristics, and the specifics of the measurement procedure.

### Data

All data were collected using the ESM methodology. Measurements consist of self-report measures of distinct emotions, with each emotion assessed using a single item. Each study also measured between-person characteristics of the participants relevant to their original study aims. In a review of the emotion time series literature detailed in Ryan et al. (2023), we identified the following seven studies which shared their data without restrictions.

The data from Rowland and Wenzel (2020) consists of measurements taken six times a day for 40 days of 125 undergraduate students from the University of Mainz in Germany. The measurements consisted of the emotion variables *happy*, *excited*, *relaxed*, *satisfied*, *angry*, *anxious*, *depressed*, and *sad*, with the items of the questionnaire phrased to query the current level of each affective state (Figure 3 indicates which items are included in which dataset). These items were scored with a visual analog slider from 0 to 100. Before the collection of the ESM measurements, participants of the study were randomly assigned to a group receiving a weekly mindfulness treatment during the ESM study and a control group.

The second dataset we use is the first dataset reported by Bringmann et al. (2016), which consists of ESM-type measurements taken 10 times a day for 7 days of 95 undergraduate students at KU Leuven in Belgium. The measures were current levels of the six emotion variables *happy*, *relaxed*, *angry*, *anxious*, *depressed*, and *disphoric*, which were scored on a continuous slider from 1 to 100. At baseline, individuals were also scored on trait neuroticism.

The data from Vrijen et al. (2018) consists of ESM-type data collected three times a day for 30 days of 138 individuals participating in a study on anhedonia in young adults (van Roekel et al., 2016). The data include measures of the level of four emotion measures *sad*, *worried*, *joy*, and *irritated* with items phrased to query the levels of these emotions since the last measurement occasion. Three measures about the experiences since the last measurement were also collected, but were excluded from our analysis due to our focus on emotions. All items were scored with a slider from 0 (*not at all*) to 100 (*very much*). The authors also assessed participants' bias toward happy faces, with low bias considered an indicator of vulnerability to depression. In their data archive, they selected the 50 individuals with the highest and lowest bias values for a between-subjects comparison.

The data from Fisher et al. (2017) consists of ESM-type data taken four times a day for approximately 30 days of 40 individuals with a generalized anxiety disorder (GAD), major depressive disorder (MDD), or comorbid GAD and MDD. The measures include ratings of the levels of 19 emotion variables or emotion-related variables since the last measurement occasion: *angry*, *worried*, *energetic*, *enthusiastic*, *content*, *irritable*, *restless*, *guilty*, *afraid*, *ahedonia*, *hopeless*, *down*, *positive*, *fatigue*, *tension*, *concentrate*, *accepted*, *threatened*, and *ruminate* and were scored on a visual analog slider from 0 (*not at all*) to 100 (*as much as possible*). In the data shared online no between-person characteristics were reported.

The data from Bringmann et al. (2013) consists of ESM-type data with 10 measurements per day over two times 6 days (with 2–3 months in-between), for 130 adults with a lifetime history of

depression and current residual depressive symptoms (Geschwind et al., 2011). The measurements were current levels of the six emotion variables *excited*, *relaxed*, *anxious*, *sad*, and *worried* and an additional variable indicating the pleasantness of the most important event since the last measurement. Since we focused on emotions in the present paper, we excluded this variable. All items were scored on a 7-point Likert scale. At baseline, individuals were also scored on trait neuroticism.

The data from Fried et al. (2020) consists of ESM-type measurements taken four times a day for 14 days of 80 undergraduate students at Leiden University in the Netherlands. The measures included ratings of nine emotion variables *relaxed*, *angry*, *worried*, *irritable*, *anhedonia*, *nervous*, *future*, *tired*, and *alone* since the last measurement occasion. We excluded the additional variables about the context and experiences of the covid pandemic. The 10 emotion variables were scored on a 5-point Likert scale. On the between-person level, the authors report a number of baseline and follow-up measurements assessing mental health and experience with COVID-19. In the analyses that follow we consider only baseline measurements of the depression anxiety stress scale (Lovibond & Lovibond, 1995).

The data of Wright et al. (2017) consists of ESM-type data with an average of 3.7 measures taken per day for 21 days of 228 individuals including individuals who were engaged in outpatient psychiatric treatment and their significant others. We report an average amount of measures taken per day because measures were initiated by individuals after social interactions in the original study (sometimes referred to as an event-contingent design; Wright et al., 2017) and not taken at pseudo-randomized intervals as in the other studies. The measures include current levels of the 31 emotion (related) variables *excited*, *angry*, *enthusiastic*, *irritable*, *guilty*, *afraid*, *nervous*, *alone*, *ashamed*, *distressed*, *hostile*, *jittery*, *scared*, *upset*, *frightened*, *shaky*, *scornful*, *disgusted*, *loathing*, *blue*, *downhearted*, *lonely*, *active*, *alert*, *attentive*, *determined*, *inspired*, *interested*, *proud*, and *strong*, which were scored on a 5-point Likert scale (from *very slightly* to *extremely*). On the between-person level, the big five personality traits were measured using the Revised NEO Personality Inventory (Costa & McCrae, 1992). In the analyses that follow we consider only measurements of neuroticism.

We provide the code to preprocess the data from the format provided by the authors into the data we used in this paper at http://github.com/jmbh/EmotionTimeSeries. We hope that these open data are useful to researchers beyond the present investigation.

## Assessing Modality and Skewness

We assess the modality and skewness of our emotion time series by classifying them as either *unimodal symmetric*, *unimodal skewed*, *bimodal*, or *multimodal* (three or more modes). One way to assign empirical distributions to those classes is to visually inspect the histograms of the empirical data and assign them to one of the classes. In Figure 1, we display the distribution of the variable *sad* for four individuals in the study of Rowland and Wenzel (2020, top row) and four individuals of the study from Bringmann et al. (2013, bottom row). While we believe that readers will sometimes differ in how they classify a given empirical distribution, we expect that most would agree with how we classified the distributions in Figure 1. If the reader agrees, then we have already established that emotion variables can take different distributional forms and that individuals can differ considerably in this respect. The goal of this work is to systematically map out the presence of these four different distributional forms and relate them to item, person, and measurement (design) characteristics.

While we consider visual inspection an appropriate—indeed, the ideal—method for this task, it has the downside that it would require us to look at 11,520 histograms in a process that is neither transparent nor reproducible. We have therefore developed a heuristic method that takes individual time series of items as input and outputs the distributional form one would choose when visually inspecting the distribution of the time series. Our method uses a standard approach of nonparametric mode estimation (for a review, see Ameijeiras-Alonso et al., 2018), which obtains a density estimate from the empirical distribution and then determines the number of modes by taking the roots of the estimated density function. After determining the number of modes in that way, we decide whether a given unimodal distribution is skewed or symmetric. We describe our heuristic method in more detail in the Appendix A.1.

The best way to validate our method is to compare it to a visual inspection. The black lines in Figure 1 show that our density estimate aligns well with the conclusion drawn by visual inspection. In case it does not, our method errs on the side of being conservative, that is, suggesting fewer modes than visual inspection would suggest. The estimates of multimodality we present in this paper are therefore a lower bound on multimodality. We present additional validation examples in the Appendix A.2, where we also show that our method outperforms Hardigan's dip statistic (Hartigan & Hartigan, 1985), the bimodality coefficient (SAS Institute Inc., 2012), a method for modality detection based on Gaussian mixture modeling (e.g., Frühwirth-Schnatter, 2006), Silverman's method (Silverman, 1981), and the excess mass-based method suggested by Ameijeiras-Alonso et al. (2019). We classify a unimodal distribution as skewed if its absolute skewness is larger than 2/3, which is again a conservative cutoff. In principle, any deviation from a skewness value of zero indicates an asymmetric distribution and since skewness is a continuous measure, any choice of the cut-off value for a binary classification will be arbitrary (see, e.g., the varying cut-offs suggested by Blanca et al., 2013; Tabachnick et al., 2007). We show in the Appendix A.3. that this cut-off leads to close to zero symmetric distributions being misclassified as skewed, while being conservative in the sense that some skewed distributions are classified as symmetric. The code to reproduce our preprocessing of the empirical data and all results in our paper can be found at http://github.com/jmbh/ModalitySkewnessPaper.

## Results

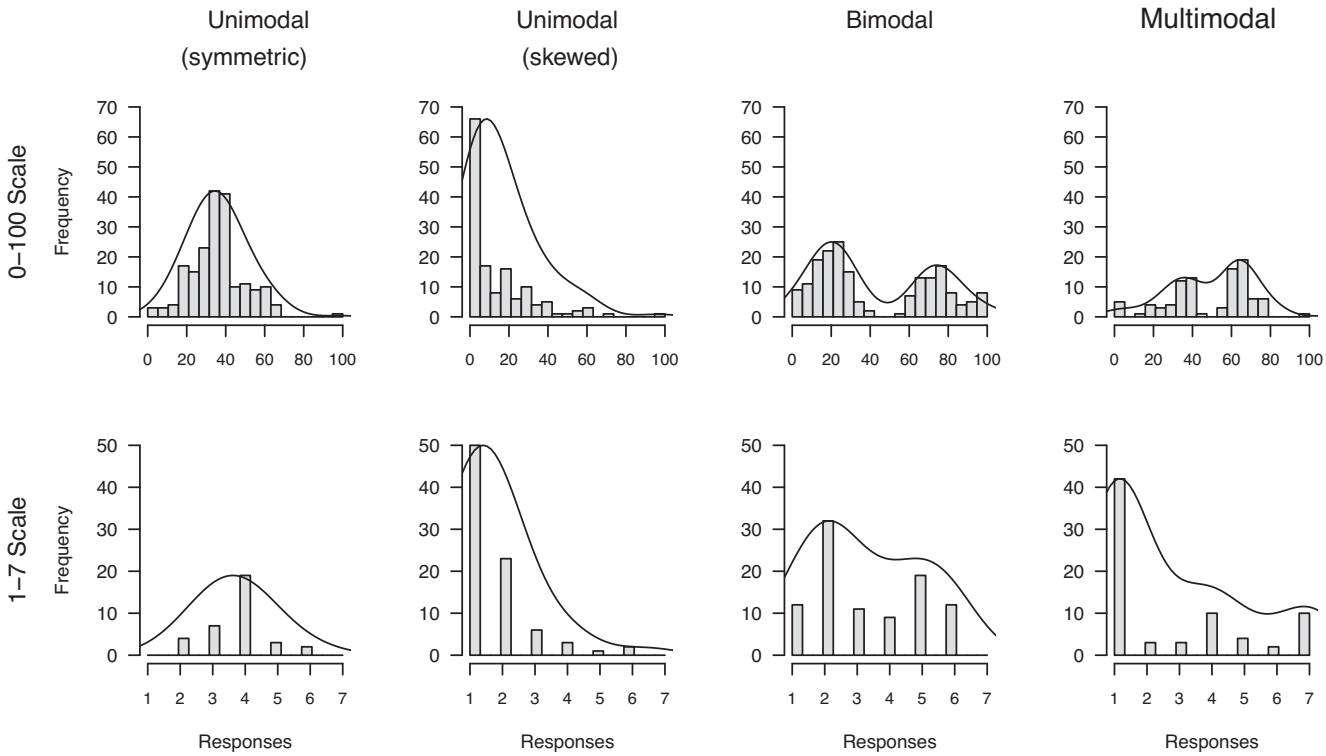
We first report our results on the number of modes in the "Multimodality" section and then analyze the skewness within the set of unimodal distributions in the "Skewness" section.

### *Multimodality*

We only considered distributions for which modality can be meaningfully assessed and therefore excluded distributions that had a standard deviation $< 0.01$ or a range of 1 (e.g., only responses for one or two categories on a Likert scale). This led to noteworthy exclusions only for the two datasets with the 1–5 Likert-scale response. Specifically, for the data of Fried et al. (2020) 1% of

**Figure 1**

*Empirical Distributions of item Sad across two studies with different response scales*




*Note.* Distributions of the item Sad measured across time in eight individuals in the study of Rowland and Wenzel (2020, top row) and the study of Bringmann et al. (2013, bottom row). The black lines indicate the density estimates of the method we use to determine the number of modes.

cases were excluded and for the data of Wright et al. (2017) 17% of cases were excluded (for details, see Appendix B).

Figure 2 displays the percentages of how many individual time series were classified as *unimodal symmetric*, *unimodal skewed*, *bimodal*, or as having *more than two modes* in each of the seven datasets. In addition, to get an overview of what the distributions look like, we displayed the density estimates for five randomly selected time series for each cell. There are two key results: first, in the majority of datasets there is a considerable proportion of multimodal distributions; and second, a considerable proportion of the unimodal distributions are skewed. In the datasets with 0–100 scales, the percentage of densities with two or more modes ranges between 21.5% and 53.3%. We see that the unimodal symmetric distributions may still be somewhat skewed, which is due to our conservative cutoff at an absolute skewness of 2/3. The skewed unimodal distributions are both left and right, but mostly right-skewed. Bimodal distributions can have two modes with about equal density and unequal densities and they can be "skewed" in the sense of having a high density at one end of the scale and a smaller density at the other end of the scale.

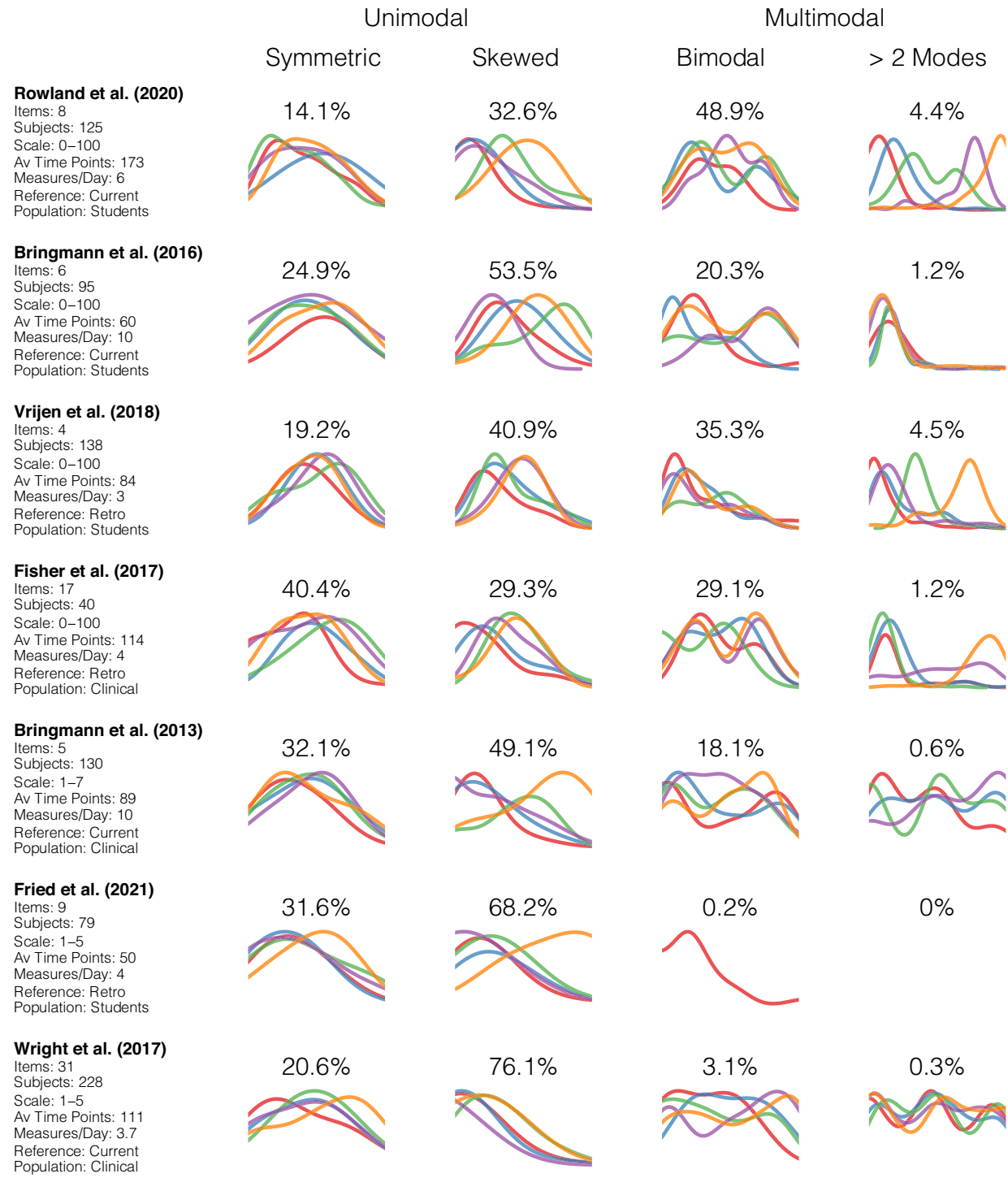
We observe much less multimodality in the datasets with Likert-scale responses. In the data with a 7-point Likert scale, we still observe 18.7% multimodality (18.1% bimodal and 0.6% with more than two modes) but for the two datasets with a 5-point Likert scale, we observe hardly any multimodality. We also see that datasets with a higher average number of time points (e.g., Rowland & Wenzel, 2020; Wright et al., 2017) exhibit slightly more multimodality. However, we do not observe an apparent relationship between the modality pattern and whether the sample was drawn from a clinical or nonclinical population.

To assess the extent of multimodality on the level of items, we average over participants separately for each dataset. We display the mean number of modes for each item and dataset in Figure 3. Given the strong influence of the measurement scale, we would not expect much consistency across items in terms of absolute modality. For example, we see that the item *happy* has a relatively high average modality of around 1.6 in the dataset of Rowland and Wenzel (2020) using a 0–100 scale, while it has a relatively low value of around 1.1 in the dataset of Bringmann et al. (2013) using a 1–5 scale.

However, we can control for the response scale by considering pairs of items that occur in a pair of datasets. For example, the items *happy* and *angry* are included in the studies of both Rowland and Wenzel (2020) and Bringmann et al. (2016). If there is consistency in the modality across items while controlling for the response scale, we would expect that one of the two items has a larger average modality than the other in both datasets. The item overlap across datasets allows for 20 such comparisons and in 17 the comparison makes sense because there is enough variation in the modality in the respective datasets. For example, in the data of Fried et al. (2020), there is no multimodality at all and therefore the comparison is not possible. In 7 of those 17 comparisons, the

**Figure 2**

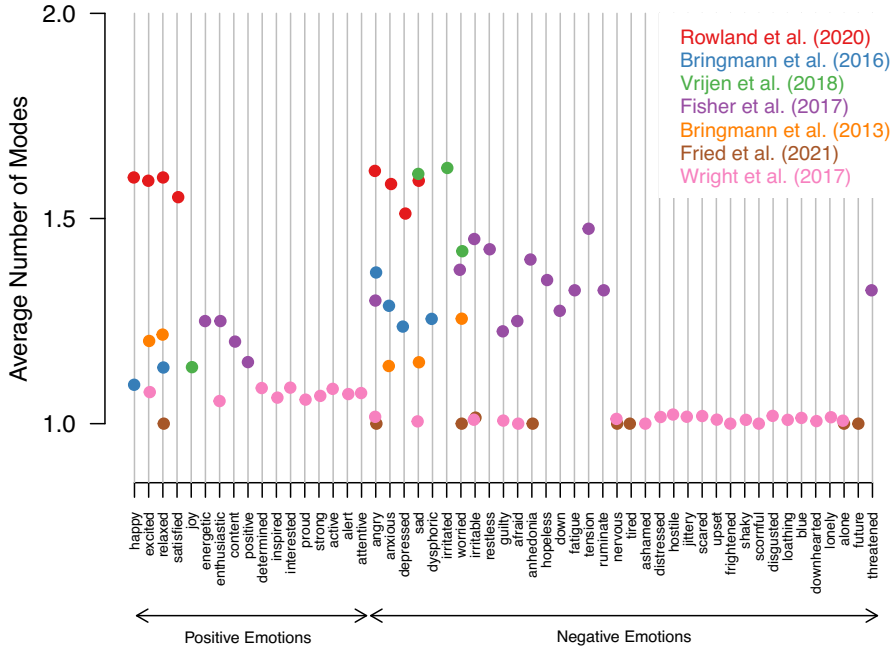
*Overview of the Prevalence of the Four Distributional Forms Within Datasets*




*Note.* The percentages are across items and individuals for each study. For each dataset and classification, we plot the density estimates of five example time series (in different colors to make them distinguishable). The modes of the empirical distributions are farther apart than in the density estimates; due to our conservative skewness cutoff, some distributions in the unimodal symmetric category are still considerably skewed. For each dataset, we also display the number of items, number of subjects, the response scale, the average number of available time points across subjects, whether the question wording was regarding the interval since the last measurement (retro) or the current moment, and the considered population. See the online article for the color version of this figure.

**Figure 3**

*Average Number of Modes for Each Item and Dataset Averaged Over Participants Within Dataset*




*Note.* The emotions are ordered from positive to negative on the *x*-axis. See the online article for the color version of this figure.

ordering of modality is consistent across the pairs of datasets. Since 50% is the consistency we expect if the modality of items is governed by chance alone, this means that we do not find consistency in modality across items.

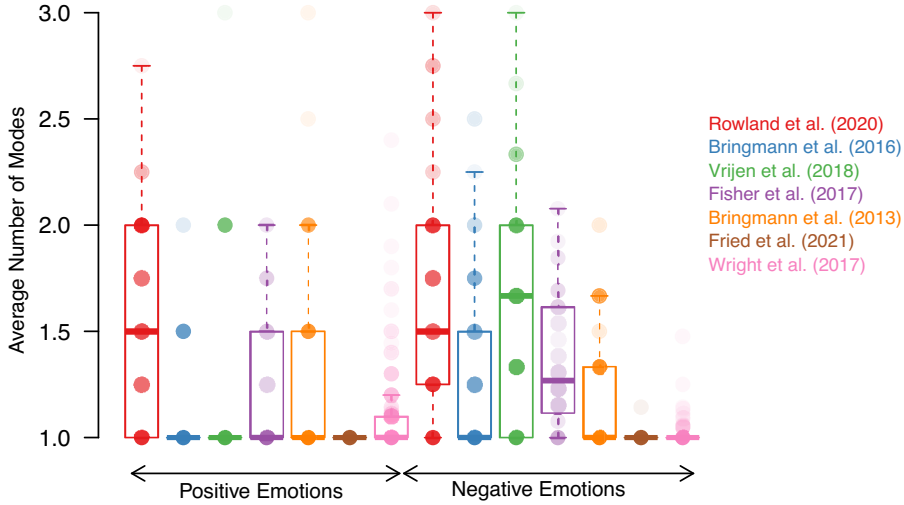
Finally, we investigate the heterogeneity in modality across individuals, averaged across items, separately for each dataset. Figure 4 displays boxplots of average modality across participants. We see that the average modality for the papers of Fried et al. (2020) and Wright et al. (2017) are very close to one, which makes sense since hardly any multimodality was detected in those datasets (see Figure 2). The datasets of Bringmann et al. (2013, 2016), and Fisher et al. (2017) also include many individuals with a very low average number of modes but also individuals with higher average modality. The datasets of Vrijen et al. (2018) and Rowland and Wenzel (2020) contain individuals that show more bimodal than unimodal responses. When looking at the three characteristics of measurement scale, average time series length, and clinical population (yes/no), we only find a clear relationship between the measurement scale and multimodality.

Moving beyond the inspection of figures, we used a three-level logistic regression model to assess which variables are associated with multimodality (1 mode vs. $\geq 2$ modes) while respecting the nested structure —items nested in participants nested in studies—of the data. We started with the simplest model that included varying intercepts for participants as well as studies, adding first item-, then participant-, and then study-level predictors if they led to an improvement in terms of the Bayesian information criterion (BIC). The person-level characteristic we considered was the time-series length for that individual, while the study-level characteristics consisted of the four design characteristics response scale, average number of time points per day, phrasing of items (as retrospective vs. current levels), and population type (as reported on the left-hand side of Figure 2). Consistent with Figure 2, the final model included a strong effect of the response scale, $\beta_{rs} = 2.92$, $SE(\beta_{rs}) = 1.38$, $p < .034$, with non-Likert type data showing more multimodality. The model also included an interaction between valence and scale, $\beta_{rs \times val} = -2.15$, $SE(\beta_{rs}) = 0.27$, $p < .001$, indicating that positive emotions that are measured on a continuous scale tended to exhibit less multimodality than negative emotions measured on a continuous scale. While the main effects of valence, $\beta_{val} = 0.36$, $SE(\beta_{rs}) = 0.20$, $p = .07$, and time-series length, $\beta_{tsl} = -0.002$, $SE(\beta_{tsl}) = 0.003$, $p = .53$, were not significant, their interaction was, $\beta_{tsl \times val} = 0.03$, $SE(\beta_{rs}) = 0.003$, $p < .001$, indicating that positive emotions tended to exhibit more multimodality the longer the time-series. The final model also included a varying slope of valence on the participant level, indicating that the effect of valence differs across participants.[1] The fixed effect of valence itself was not significant, $\beta_{val} = 0.36$, $SE(\beta_{rs}) = 0.20$, $p = .07$. The number of measurements per day, the population (students vs. clinical), and whether the design was retrospective or momentary were not included in the final model. The details of this analysis can be found in Appendix D.

[1] An intermediate model also included a varying slope at the study level, but this model was outperformed by a model with the valence and scale interaction. Including both the varying slope on the study level and the interaction led to a design matrix that was not positively definite, so including both was not possible.

**Figure 4**

*The Variation in Modality Across Individuals and Datasets Averaged Across Positive and Negative Emotions*




*Note.* See the online article for the color version of this figure.

As reported above, most datasets included person-level characteristics, and we studied whether those are associated with multimodality in each study separately. For this purpose, we fit a two-level model for each study separately with varying intercepts and varying effects of valence and examined whether adding the available person-level variable improved model fit. For Bringmann et al. (2013, 2016), and Wright et al. (2017), the neuroticism variable failed to improve model fit. Similarly, the person-level predictors' mindfulness treatment versus control and bias toward happy faces (low vs. high) in study Rowland and Wenzel (2020) and Vrijen et al. (2018), respectively, both did not exhibit a significant association with multimodality. We could not run the model for study Fried et al. (2020) separately, because only two individuals showed multimodality, leaving almost no variance to explain.

### *Skewness*

We next consider how skewness in unimodal distributions is related to item-, participant-, and study-characteristics. For this analysis, we do not require a cutoff and we therefore use the continuous values of skewness. Figure 5 displays mean skewness across participants for each item separately for the seven datasets.

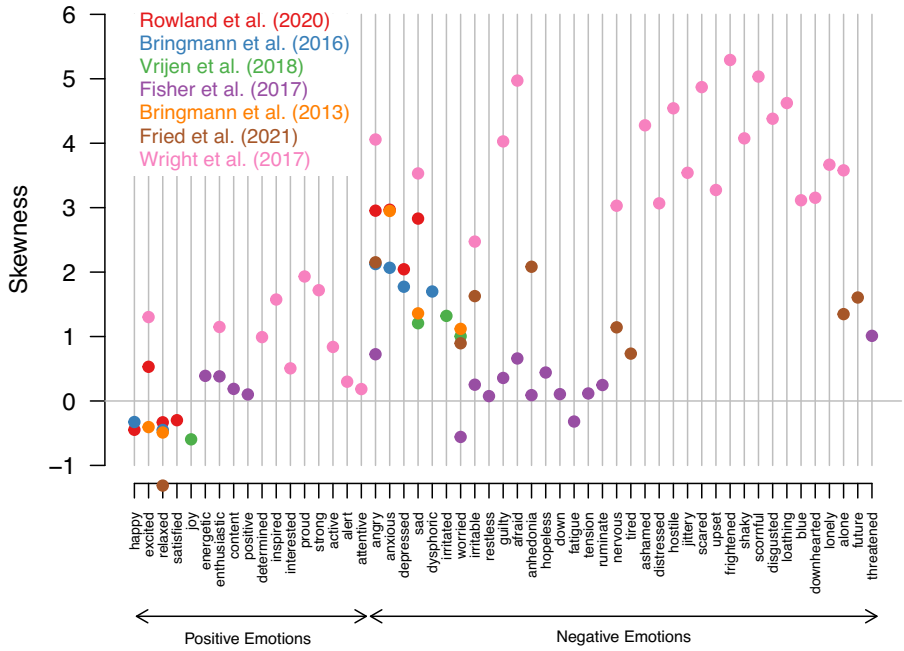
We see that positive emotions tend to be negatively skewed and also exhibit less skew in absolute value than negative emotions. However, we also see that there is a considerable amount of heterogeneity across items. For example, the positive emotion *excited* has been measured in three studies and has a mean skewness ranging from $-0.4$ to 1.4. In contrast, the negative emotion *anxious* has also been measured in three studies and its skewness is ranging from 2.1 to 3.1. However, the inconsistency in positive/negative skewness in positive emotions seems to be largely explained by the type of scale, since we only observe positive skew for positive emotions for the datasets using a 1–5 Likert scale.

Figure 6 shows the variation in skewness across individuals per study, averaged across items of positive and negative valence, respectively. Consistent with the results in Figure 5, we see that positive emotions are typically negatively skewed, negative emotions are typically positively skewed, and the absolute skewness of positive emotions is much smaller than that of negative emotions for all studies except for Fried et al. (2020) and Wright et al. (2017), which used the 1–5 Likert scale. We also see that there is considerable heterogeneity across individuals and that the amount of heterogeneity depends on valence (less for positive emotions) and studies (e.g, less in Fisher et al., 2017 than in Rowland & Wenzel, 2020).

Next, we used a multilevel model to quantify and explain heterogeneity at the item, subject, and study level. To ease the interpretation of parameters, we use the absolute value of skewness as the dependent variable, such that zero represents an unskewed variable, and positive numbers represent a high degree of either positive or negative skewness. Similar to the analysis of modality reported above, we used a multilevel linear regression model to assess which variables are associated with skewness while respecting the nested structure—items nested in participants nested in studies—of the data. We start with the simplest model that includes varying intercepts for participants as well as studies, adding first item-, then participant-, and then study-level predictors if they led to an improvement in terms of the BIC. The simplest model, containing only a varying intercept for participants and studies, yielded an intraclass correlation coefficient (ICC) of 0.188 at the participant level and 0.085 at the study level. This means that around 27.3% of the overall variance is attributable to participant and study variation, with more than twice as much on the participant level than the study level. This observation is also born out by Figure 6, where we see that across studies the mean skewness is relatively similar while the spread of skewness per participant within a study is much more extreme. The ICC tells us that the average variation

**Figure 5**

*Mean Skewness of Unimodal Distributions for Each Item Separately for Each Dataset Averaged Over Participants Within Each Dataset*



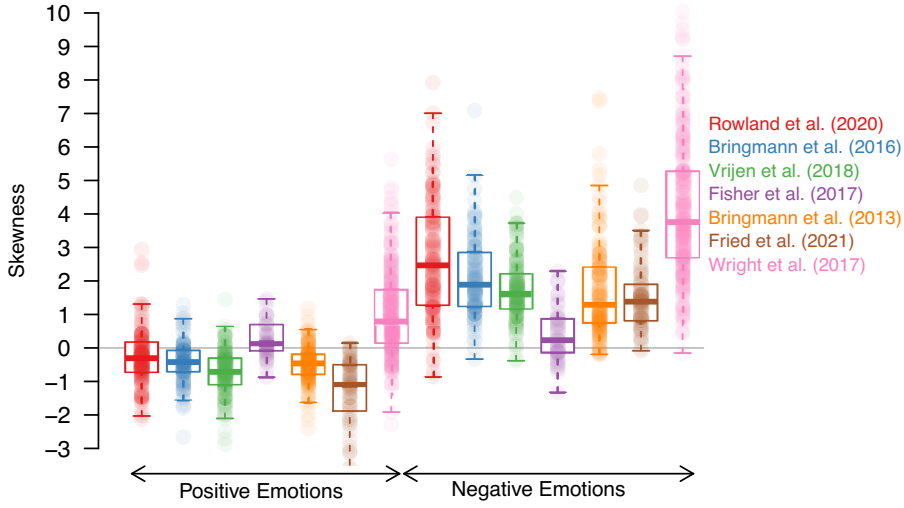

*Note.* The emotions are ordered from positive to negative on the *x*-axis. See the online article for the color version of this figure.

within a participant is even greater than the variation across participants, accounting for 76.4% of the overall variance.

In our model, we dummy code the valence of items (val) such that zero represents negative items and one positive item. The final model included a positive intercept, $\beta_0 = 2.11$, $D(\beta_{rs})$ $S = 0.383$, $p = .001$, and a strong effect of valence, $\beta_{val} = -1.27$, $SD(\beta_{val}) = 0.295$, $p = .004$. Together, these parameters indicate that negatively valenced items were on average highly skewed and that positively

**Figure 6**

*The Variation in Raw Skewness Across Individuals and Datasets Averaged Across Items, Separately for Positive and Negative Emotions*




*Note.* See the online article for the color version of this figure.

valenced items were significantly less skewed than negative items. The model also included a main effect of time-series length, $\beta_{tsl} = 0.014$, $SD(\beta_{rs}) = 0.002$, $p < .001$, indicating that longer time-series tended to exhibit more skew for negatively valenced items. There was also an interaction between valence and time-series length, $\beta_{val \times tsl} = -0.012$, $SD(\beta_{val \times tsl}) = 0.002$, $p < .001$, indicating that the effect of increasing time series-length was significantly less strong for positive compared to negative items. The final model did not include any study-level predictors; further details of our analysis can be found in Appendix E.

Focusing on analyzing the relationship of between-person characteristics with skewness in each specific study, we found a relationship between neuroticism and skewness. For Bringmann et al. (2013, 2016), and Wright et al. (2017), above-average neuroticism was associated with lower skewness for negatively valenced items with a much smaller effect for the skew of positively valenced items. For Bringmann et al. (2013) and Wright et al. (2017), the model with neuroticism outperformed the model without Level-2 predictors according to the Akaike information criterion (AIC), BIC, and $\chi^2$ test. See the Appendix E.1. for parameter estimates. In the other studies, there was no clear effect of the reported between-person characteristics on skewness: adding the Level-2 predictors failed to improve model fit according to the BIC in each case.

## Discussion

In this paper, we systematically assessed the modality and skewness of emotion measurements across seven datasets, comprising 11,520 univariate time series spread over 835 individuals and 53 unique items. Each study differed in a variety of aspects, including the used response scale, the employed sampling scheme, the items, the sampled popultion, and the considered between-person characteristics. To assess modality, we used a heuristic method developed and validated for the current investigation, implemented in R and accessible from https://github.com/jmbh/ModalitySkewnessPaper. Our validation suggests that the heuristic method is conservative but performs well compared to more general alternatives in the seven datasets considered here (see the Appendix A.2.). Researchers who wish to analyze the modality of their own data may make use of our heuristic method, but should note that we validated it only for the types of data and scales analyzed in the present paper.

### Discussion of Results

Our main result is that multimodality is highly prevalent in emotion measurements with up to 48% of distributions within datasets showing two modes or more. In addition, we found that a large proportion (up to 68%) of unimodal distributions were considerably skewed. In only one out of the seven datasets considered (Fisher et al., 2017) was a unimodal symmetric distribution the most common classification. Since we designed our classification method to be conservative, the percentages of multimodality and skewness should be interpreted as lower bounds. If the reader were to go through the empirical histograms themselves, they would likely classify many more distributions as multimodal and many more unimodal distributions as skewed. In addition to these aggregate results, we also explored how modality and skewness relate to measurement, item type, and the sampled population.

We considered characteristics on the level of items, individuals, and datasets and explored whether they could explain the prevalence of multimodality. The effect that clearly stood out was that multimodality strongly depends on the response scale. The prevalence of multimodality was high for the 0–100 scales (between 21.5% and 53.3%), much lower for the 1–7 scale (18.2%), and extremely low for the 1–5 scale (0.2% and 3.8%). There are two explanations for this finding. The first one is that individuals indeed experience emotion intensity in a multimodal way, but this multimodality is largely masked when measuring intensity using only a few ordinal categories. This is supported by the fact that we still have sizeable multimodality in the study that used a 7-point Likert scale, but almost none in the two studies that used a 5-point Likert scale. In Appendix C, we demonstrate how thresholding a multimodal distribution into a Likert scale can mask multimodality. The alternative explanation is that multimodality is somehow induced by the 0–100 slider scale. For example, if the slider is initialized in the middle of the scale at 50 and individuals are required to move, they might find it difficult to move back to exactly the middle, especially on a relatively small smartphone screen. The initialization at 50 might therefore induce bimodality. From correspondence with the authors, we know that this was indeed the case in the procedure of Rowland and Wenzel (2020) (48.9% bimodal) and Fisher et al. (2017) (29.1% bimodal), but not in (Bringmann et al., 2016) (20.5% bimodal) and (Vrijen et al., 2018) (35.3% bimodal). These results suggest that initializing the slider at 50 may induce bimodality to some extent, but they also show that bimodality is not explained away by this design choice. The high prevalence of multimodality has major consequences for theory, measurement, and time series modeling, which we discuss in turn below.

Turning to skewness, our most pronounced result is that negative emotions exhibit much higher skewness than positive emotions, a finding that held across studies. One explanation for this result is that the intensity of negative emotions tends to be lower than for positive emotions in the general population (e.g., Diener & Diener, 1996; Zelenski & Larsen, 2000), which raises the question of whether skewness provides additional information beyond the location of the distribution. We investigated this in Appendix F, where we predicted skewness by the proportion of cases in the lowest 10% quantile (0–100 scales) or the lowest ordinal category (Likert scales). This analysis shows that the skewness and the location of the distributions are correlated, but that skewness provides additional information beyond location. Similarly, positive emotions tended to exhibit negative skewness, indicating that measurements cluster at the top of the scale, with less frequent lower responses. Across three studies, we found that participants who had high neuroticism, a trait typically associated with poor emotion regulation (Bringmann et al., 2013, 2016; Wright et al., 2017), exhibited less skewed responses to negative emotion items. We did not detect any strong study-level predictors of skewness, though with only seven studies we had low statistical power to detect these effects. An exploratory analysis indicated that retrospective phrasing of items (rating emotion levels since the last measurement occasion in contrast to current levels) may be predictive of lower absolute skewness for negatively valenced items, although no strong statements can be made about this effect (see Appendix E).

Other study-level characteristics not considered here may be of interest to future research. For example, Wright et al. (2017) used an event-contingent design, with participants prompted to fill out surveys immediately following social interactions, while all other studies used a signal-contingent design, with participants prompted to fill out surveys at (randomized) occasions throughout the day. Studying the effect of these types of design choices would either require access to many more empirical datasets with a greater variety of design choices than here, or systematic variation of one design choice at a time in a new empirical study. In either case, further research is needed to disentangle the effect of different design choices on response distributions of emotion items.

For both modality and skewness, we found an interaction between the number of observations in a time series for an individual and the valence of an item: Longer time series were associated with multimodality for positive items, and higher skewness for negative items. The length of the time series is determined both by the sampling frequency and by the observation window, both of which vary across the seven studies considered. Studies that cover a longer observation window (e.g., 30 days for Vrijen et al., 2018 vs. 7 days for Bringmann et al., 2016) also typically took less frequent measurements per day (e.g., 3 times a day for Vrijen et al., 2018 vs. 10 times a day for Bringmann et al., 2016). As such, we are not able to disentangle to what degree these effects are due to the frequency of measurements or the timescale of the study. Furthermore, while for simplicity, we refer throughout to measurements of emotion, some readers may interpret certain items or study designs as more closely mapping to affect or mood, closely related constructs, which are typically considered to evolve over a slower timescale (Frijda et al., 1991; Kuppens et al., 2022). The distinction between modality and skewness of emotion versus mood items was not possible in our analysis.

## Theories of Emotion Dynamics

The high prevalence of multimodality has important implications for theorizing, because it provides information about how emotions are experienced in daily life. Symmetric unimodal emotion measures would imply that individuals experience a given emotion with a certain typical intensity and vary to some extent from that typical intensity as a reaction to their environment. In contrast, the multimodal distributions we identified in this paper imply that individuals experience emotions in qualitatively different states and vary around the typical intensity of a given state. For many emotions this is very plausible: for example, most people are generally not angry, but occasionally they experience a situation that makes them switch into a state of anger for a certain time interval. After the situation is resolved, the person switches back into a state of no anger. It could also be possible that people are generally not angry, but switch to different alternative states, depending on how much anger the given situation elicits. The alternative states might overlap and thereby produce the skewed distributions we identified. The high prevalence of multimodality observed in our data suggests that it is a stable phenomenon of how people experience emotions and therefore needs to be accounted for by any (formal) theory of emotion dynamics.

We showed that there is considerable heterogeneity in both modality and skewness across individuals. A promising avenue for future research would be to explore to what extent this heterogeneity can be explained by inter-individual differences in other variables. In the present paper, we explored such analyses with the measures that were available in the seven datasets and showed that individuals with higher neuroticism have on average a more negative skew for negative emotions. In future work, one could relate interindividual differences in modality and skewness to symptom patterns of mental disorders, which might explain variations in emotion dynamics and regulation, and in turn help shed light on the etiology of the disorder (Lincoln et al., 2022).

One interesting extension of our analysis of distributional forms would be to consider *multivariate* distributions. A limitation of our univariate analysis is that it does not allow us to assess the modality of the joint distribution of, for example, positive and negative emotions. We could observe a bimodal distribution for both, but these bimodal marginal distributions are consistent with many bivariate distributions which have very different theoretical interpretations. For example, we do not know whether the data points fall on a straight line, indicating a linear relationship; or whether they occur only in certain "quadrants," for example, with many observations in "low positive, low negative," but almost no observations in "high positive, high negative" (e.g., Loossens et al., 2020).

Another promising direction would be to explore features of emotion time series in the time domain. In our analysis, we did not use the fact that measurements are ordered in time and therefore ignored a considerable amount of information in the data. Taking time into account one could for example determine how many times individuals switch between states, and how long they stay in each state. A related analysis is the study of intensity profiles of emotions (e.g., Résibois et al., 2018; Verduyn et al., 2012), which looks into how the intensity of an emotion unfolds in a given emotional episode. In this context, researchers have also found bimodal distributions across time, with an initial peak in emotion intensity, followed by a reduction, and a second peak. However, this type of modality is likely to be different from the one observed in this paper, because the time between measurements in the studies considered here is likely too long to capture a single emotional episode.

## Emotion Measurements

A key result of our investigation is the much more pronounced multimodality in studies using a 0–100 continuous slider scale compared to studies using Likert scales. This might be explained by the fact that Likert scales can mask multimodality (see Appendix C); for example, to arrive at a multimodal distribution with five response types, Options 1, 3, and 5 must have high peaks while Options 2 and 4 must have low peaks—a pattern that strongly constrains the data. In general, for the data to show $K$ modes the response scale must have at least $2K - 1$ response options. Additionally, while our results suggest that multimodality can be induced by a measurement procedure that initializes the slider at, for example, the middle of the scale, we also saw that multimodality is not explained away by this artifact. The remaining variation across studies may partly be explained by measurement properties beyond the response scale (see, e.g., Brose et al., 2020; Trull & Ebner-Priemer, 2020). It is, of course, possible that multimodality is induced by using a 0–100 scale, however, if that is the case, it is unclear to us what mechanism would produce such a phenomenon. As such, the extent to which different measurement procedures can induce multimodality is an important area for future research. Based on the analysis outlined above and in Appendix C, we would suggest that, if researchers

believe a multimodal distribution underlies their data, then a 0–100 scale should be preferred.

In our analysis, we had to exclude a number of cases due to extremely little variance or the endorsement of only two response categories in the two datasets with the 1–5 scale. By definition these cases could be considered uni- and bimodal, respectively, however, we chose to exclude these because typically these response categories are treated as continuous, and so we wished to capture only multimodality in responses which exhibit a reasonable degree of variation across the measurement scale. In the case where extremely little variance was found, the point mass tended to cluster around the ends of the rating scale. Although this point mass at the end of scales was not the focus of the present paper, we consider it an extremely interesting direction for future work (see also Ram et al., 2017). Another interesting avenue for future research would be to investigate the impact of certain details of the measurement process on the response distributions. For example, when using a 0–100 scale, using ticks at, say, 0, 50, and 100, might create a response style that artificially increases the response frequency of these categories (Matejka et al., 2016).

It seems intuitively plausible that multimodality should increase with the observation period. For example, while I may not be sad during a study that takes a week, I may well exhibit sadness during a study that takes a month. One possible explanation could be that most events in daily life are relatively common and elicit no or very small emotion intensities. Events that cause stronger emotion intensities are less frequent and might therefore not be captured in shorter time series, thus leading to less multimodality in emotion intensities in shorter time series. However, this likely is not the full explanation for why we observe multimodality in some time series but not in others; in many cases, bimodality is clearly produced by frequent switching between high and low response categories (for an example, see Figure F2 in Appendix F). In the current paper, we did not include the study duration as a three-level predictor because it did not vary substantially across studies. Instead, we used the average number of time points, but this confounds the sampling frequency with the study duration. Future research that has access to a larger set of studies may wish to test the relationship between multimodality and study duration explicitly.

The implications of these results for measurement depend on one's beliefs about the nature of the emotion variable one is trying to measure. If one believes that emotions are inherently continuous, then this result indicates that 0–100 response scales allow one to capture more granular features of emotion responses such as multimodality than 5- or 7-point Likert scales. However, one might interpret the frequency of uni- and bimodal response patterns in these data as an indication that emotion responses consist most meaningfully of one or two discrete states, and that variation around these modes may be indicative of processes such as measurement error. From this perspective, one might conclude that reducing the number of response categories to two or three might capture most of the interesting information to be found in responses to emotion questionnaires in daily life. The benefit of having fewer response categories would be that statistical analysis of this data is considerably more straightforward than with 5- or 7-point scales, which are typically treated as continuous for pragmatic reasons. With fewer response categories, the dynamics which govern movement from one category to another can be modeled directly, for instance using Markov models (Jackson, 2007).

## Time Series Modeling

The high prevalence of multimodal distributions and skewness has important consequences for statistical modeling. Currently, the most popular way to analyze emotion time series is the VAR model (e.g., Vanhasbroeck et al., 2021). However, it is well-known that the VAR model has only a single equilibrium to which it returns after perturbations drawn from a Gaussian distribution (Hamilton, 1994). This implies that the VAR model gives rise to unimodal symmetric distributions. Consequently, the VAR model will fit emotion time series that exhibit multimodality or skewness very poorly. One can easily check this by generating data from the estimated VAR model and comparing these data to the empirical data (we demonstrate this for a single individual from the data of Rowland and Wenzel (2020) in Appendix G). In addition, interpreting its parameters may be misleading (for a discussion of this see Haslbeck & Ryan, 2022). VAR models may still be a useful way to characterize a multivariate time series. However, the presence of multimodality and skewness implies that the VAR model will fit the data poorly, that it is not a plausible generating mechanism, and consequently, that the interpretation of its parameters is not straightforward.

It may therefore be more appropriate to use models that are able to capture different states and the transitions between them. One prominent avenue may be HMMs, which allow the specification of different states and capture the probability of switching between states in a transition matrix (e.g., Catarino et al., 2020; Cochran et al., 2016; Dempsey et al., 2017; Hamaker et al., 2016; Prisciandaro et al., 2019; Visser & Speekenbrink, 2010; Yee et al., 2021). HMMs can reproduce multimodality and skewness in empirical data (Ryan et al., 2023). Another, often neglected avenue would be to focus on descriptive analyses instead of using advanced statistical models. This can include looking at means, variances, covariances (see, e.g., Zelenski & Larsen, 2000, for an excellent demonstration), and distributional forms as done in this paper. Using a range of exploratory data analysis techniques can help us get a better understanding of our data and help us to choose appropriate statistical methods.

## Conclusion

In this paper, we systematically mapped out the modality and skewness of emotion measurements in psychological time series datasets and found a high prevalence of multimodality and skewness. We discussed the implications of our findings for theory development, measurement practices, and statistical modeling. We hope that our investigation can help advance the study of emotion dynamics.

## References

Ameijeiras-Alonso, J., Crujeiras, R. M., & Rodríguez-Casal, A. (2018). *Multimode: An R package for mode assessment.* Preprint arXiv. arXiv:1803.00472.

Ameijeiras-Alonso, J., Crujeiras, R. M., & Rodríguez-Casal, A. (2019). Mode testing, critical bandwidth and excess mass. *TEST*, *28*(3), 900–919. https://doi.org/10.1007/s11749-018-0611-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, *7*(1), 108–116. https://doi.org/10.1162/neco.1995.7.1.108

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*(4), 425–435. https://doi.org/10.1177/1073191116645909

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4), Article e60188. https://doi.org/10.1371/journal.pone.0060188

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, *20*(4), 677–699. https://doi.org/10.1037/emo0000583

Catarino, A., Fawcett, J. M., Ewbank, M. P., Bateup, S., Cummins, R., Tablan, V., & Blackwell, A. D. (2020). Refining our understanding of depressive states and state transitions in response to cognitive behavioural therapy using latent Markov modelling. *Psychological Medicine*, *52*(2), 332–341. https://doi.org/10.1017/S0033291720002032

Cochran, A., McInnis, M., & Forger, D. (2016). Data-driven classification of bipolar I disorder from longitudinal course of mood. *Translational Psychiatry*, *6*(10), e912–e912. https://doi.org/10.1038/tp.2016.166

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological Assessment*, *4*(1), 5–13. https://doi.org/10.1037/1040-3590.4.1.5

Davidson, R. J. (1998). Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, *12*(3), 307–330. https://doi.org/10.1080/026999398379628

Dempsey, W. H., Moreno, A., Scott, C. K., Dennis, M. L., Gustafson, D. H., Murphy, S. A., & Rehg, J. M. (2017). iSurvive: An interpretable, event-time prediction model for mHealth. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 970–979). http://proceedings.mlr.press/v70/dempsey17a.html

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological Science*, *7*(3), 181–185. https://doi.org/10.1111/j.1467-9280.1996.tb00354.x

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology*, *126*(8), 1044–1056. https://doi.org/10.1037/abn0000311

Fried, E. I., Papanikolaou, F., & Epskamp, S. (2020). Mental health and social contact during the COVID-19 pandemic: An ecological momentary assessment study. *Clinical Psychological Science*, *10*(2), 340–354. https://doi.org/10.1177/21677026211017839

Frijda, N. H. (2017). *The laws of emotion*. Psychology Press.

Frijda, N. H., Mesquita, B., Sonnemans, J., & Van Goozen, S. (1991). The duration of affective phenomena or emotions, sentiments and passions. K. T. Strongman (Ed.), *International review of studies on emotion* (Vol. 1., pp. 187–225). Wiley.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer.

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628. https://doi.org/10.1037/a0024595

Hamaker, E. L., Ceulemans, E., Grasman, R., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment*, *23*(4), 436–446. https://doi.org/10.1177/1073191116632339

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton University Press.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84. https://doi.org/10.1214/aos/1176346577

Haslbeck, J. M., & Ryan, O. (2022). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, *57*(5), 735–766. https://doi.org/10.1080/00273171.2021.1896353

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). *Multimodality and skewness in emotion time series*. https://doi.org/10.31234/osf.io/qudr6

Jackson, C. (2007). *Multi-state modelling with R: The MSM package*. MRC Biostatistics Unit. https://cran.r-project.org/web/packages/msm/vignettes/msm-manual.pdf

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., Dejonckheere, E., Kalokerinos, E., & Koval, P. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science*, *3*(2), 505–515 https://doi.org/10.31234/osf.io/y4aqh

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Lebo, M. A., & Nesselroade, J. R. (1978). Intraindividual differences dimensions of mood change during pregnancy identified in five p-technique factor analyses. *Journal of Research in Personality*, *12*(2), 205–224. https://doi.org/10.1016/0092-6566(78)90098-3

Lincoln, T. M., Schulze, L., & Renneberg, B. (2022). The role of emotion regulation in the characterization, development and treatment of psychopathology. *Nature Reviews Psychology*, *16*, 272–286. https://www.nature.com/articles/s44159-022-00040-4

Loossens, T., Mestdagh, M., Dejonckheere, E., Kuppens, P., Tuerlinckx, F., & Verdonck, S. (2020). The affective Ising model: A computational account of human affect dynamics. *PLoS Computational Biology*, *16*(5), Article e1007860. https://doi.org/10.1371/journal.pcbi.1007860

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5421–5432). https://doi.org/10.1145/2858036.2858063

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. (2013). Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology*, *4*, Article 700. https://doi.org/10.3389/fpsyg.2013.00700

Prisciandaro, J. J., Tolliver, B. K., & DeSantis, S. M. (2019). Identification and initial validation of empirically derived bipolar symptom states from a large longitudinal dataset: An application of hidden Markov modeling to the systematic treatment enhancement program for bipolar disorder (STEP-BD) study. *Psychological Medicine*, *49*(7), 1102–1108. https://doi.org/10.1017/S0033291718002143

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

Résibois, M., Kalokerinos, E. K., Verleysen, G., Kuppens, P., Van Mechelen, I., Fossati, P., & Verduyn, P. (2018). The relation between rumination and temporal features of emotion intensity. *Cognition and Emotion*, *32*(2), 259–274. https://doi.org/10.1080/02699931.2017.1298993

Rowland, Z., & Wenzel, M. (2020). Mindfulness and affect-network density: Does mindfulness facilitate disengagement from affective experiences in daily life? *Mindfulness*, *11*(5), 1253–1266. https://doi.org/10.1007/s12671-020-01335-4

Ryan, O., Dablander, F., & Haslbeck, J. M. B. (2023). Towards a generative model for emotion dynamics. https://psyarxiv.com/x52ns/

SAS Institute Inc. (2012). *Stat 12.1: User's guide*.

Schimmack, U., Oishi, S., Diener, E., & Suh, E. (2000). Facets of affective experiences: A framework for investigations of trait affect. *Personality and Social Psychology Bulletin*, *26*(6), 655–668. https://doi.org/10.1177/0146167200268002

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1991.tb01857.x

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(1), 97–99. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1981.tb01155.x

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Pearson.

Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PLoS ONE*, *10*(12), Article e0145450. https://doi.org/10.1371/journal.pone.0145450

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

van Roekel, E., Masselink, M., Vrijen, C., Heininga, V. E., Bak, T., Nederhof, E., & Oldehinkel, A. J. (2016). Study protocol for a randomized controlled trial to explore the effects of personalized lifestyle advices and tandem skydives on pleasure in anhedonic young adults. *BMC Psychiatry*, *16*(1), Article 182. https://doi.org/10.1186/s12888-016-0880-z

Vanhasbroeck, N., Ariens, S., Tuerlinckx, F., & Loossens, T. (2021). Computational models for affect dynamics. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 213–260). Springer.

Verduyn, P., Mechelen, I. V., & Frederix, E. (2012). Determinants of the shape of emotion intensity profiles. *Cognition and Emotion*, *26*(8), 1486–1495. https://doi.org/10.1080/02699931.2012.662152

Verduyn, P., Van Mechelen, I., & Tuerlinckx, F. (2011). The relation between event processing and the duration of emotional experience. *Emotion*, *11*(1), 20–28. https://doi.org/10.1037/a0021239

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21. https://doi.org/10.18637/jss.v036.i07

Vrijen, C., Hartman, C. A., Van Roekel, E., De Jonge, P., & Oldehinkel, A. J. (2018). Spread the joy: How high and low bias for happy facial emotions translate into different daily life affect dynamics. *Complexity*, *2018*, Article 2674523. https://doi.org/10.1155/2018/2674523

Wichers, M., Wigman, J., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362–367. https://doi.org/10.1177/1754073915590623

Wright, A. G., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology*, *126*(7), 898–910. https://doi.org/10.1037/abn0000286

Yee, M. A., Yocum, A. K., McInnis, M. G., & Cochran, A. L. (2021). Dynamics of data-driven microstates in bipolar disorder. *Journal of Psychiatric Research*, *141*, 370–377. https://doi.org/10.1016/j.jpsychires.2021.07.021

Zelenski, J. M., & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality*, *34*(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275

(*Appendices follow*)

# Appendix A

## Method for Determining the Distributional Form

We developed our heuristic method for detecting the distributional form through an iterative process in which we tuned the parameters of our method to best align with what we would conclude based on visual inspection of the empirical data. This visual inspection is, after all, the gold standard for determining multimodality. We do not, however, suggest our method as a general tool to detect multimodality in empirical data more broadly, because we have tuned it to the specific datasets at hand. We describe our method in more detail in the Appendix A.1. and provide a visual and simulation-based validation for the modality estimation as well as skewness in the Appendix A.2. and Appendix A.3., respectively.

### Appendix A.1. Description of the Method

Our method is based on density estimation. In the first step, we add Gaussian noise with standard deviation $\sigma$ to each measurement, which acts as a form of smoothing (e.g., Bishop, 1995). We set $\sigma$ depending on the range of the measurements. In particular, for the data with 100 response categories, we set $\sigma = (0.035 + (Q/5)) \times 100$, where $Q$ is the number of observations with the value of the mode divided by the number of data points. For example, if the data would consist of {10,10,10,30,20,20}, then $Q = 3/6$, since the value at the mode (which is 10) appears three times. For the Likert scale data, we simply set the standard deviation of the noise to $Q$. In the second step, we estimate the density of these noise-augmented data with a Gaussian kernel using the method by Sheather and Jones (1991), which selects the bandwidth using a data-driven approach, which we scaled by a factor of 2. Note that if we had not added Gaussian noise, the density estimation method would estimate a mode for every category in the Likert-scale data, which is clearly undesirable. Making in general, making $\sigma$ dependent on $Q$ introduces a penalty for distributions in which almost all responses have the same value (typically zero), and the remaining values are scattered across the rest of the scale. Lastly, to obtain the number of modes we take the derivative of the density estimate and count its roots. In order to avoid that our results are dependent on specific noise draws we repeat the procedure 10 times for each set of measurements and pick the number of modes that have been estimated most often. In case of a tie, we select the lower modality. For unimodal distributions, we further calculate their empirical skewness.

### Appendix A.2. Validation of Modality Estimation

We validate our method with two strategies. The more important one is the first, in which we look at empirical distributions and see whether the modality estimated by the method lines up with what we would decide based on visual inspection. Second, we simulate from unimodal and multimodal distributions and see whether our method correctly estimates the modes. We consider this strategy less important, because the data-generating distributions we use are likely not representative of the data-generating mechanisms of empirical data. However, they do serve as a sanity-check, because our method should perform well in these idealized scenarios. We also compare our method with five other popular methods for modality detection: Hardigan's dip statistic (Hartigan & Hartigan, 1985), the bimodality coefficient (SAS Institute Inc., 2012), Gaussian mixture modeling (Frühwirth-Schnatter, 2006), Silverman's method (Silverman, 1981), and the excess mass-based method suggested by Ameijeiras-Alonso et al. (2019).

#### *Appendix A.2.1. Validation With Empirical Examples*

Figure A1 displays nine empirical distributions taken from the data of Rowland and Wenzel (2020) that illustrate the performance of our method in detecting the number of modes in continuous data. Panel (a) seems to be clearly unimodal and our method correctly estimates a single mode. Panels (b) and (c) could be seen as a skewed unimodal distribution, but they could also be seen as a bimodal/multimodal distribution with a large density at zero and additional smaller modes. However, the modes are not very well separated, and our method estimates one mode. Panels (d), (e), and (f) show multimodal distributions and our method correctly identifies two modes in all cases. However, in panel (d) we see that the density estimate is close to unimodal, which is due to the fact that the distance between the two modes is relatively small. Panels (g), (h), and (i) display very clear multimodal distributions which are consequently also picked up by the method. These examples show that our method detects multimodality if it is clearly pronounced, but estimates unimodality in cases that are unimodal or that are not clearly multimodal. Our method is therefore quite conservative.

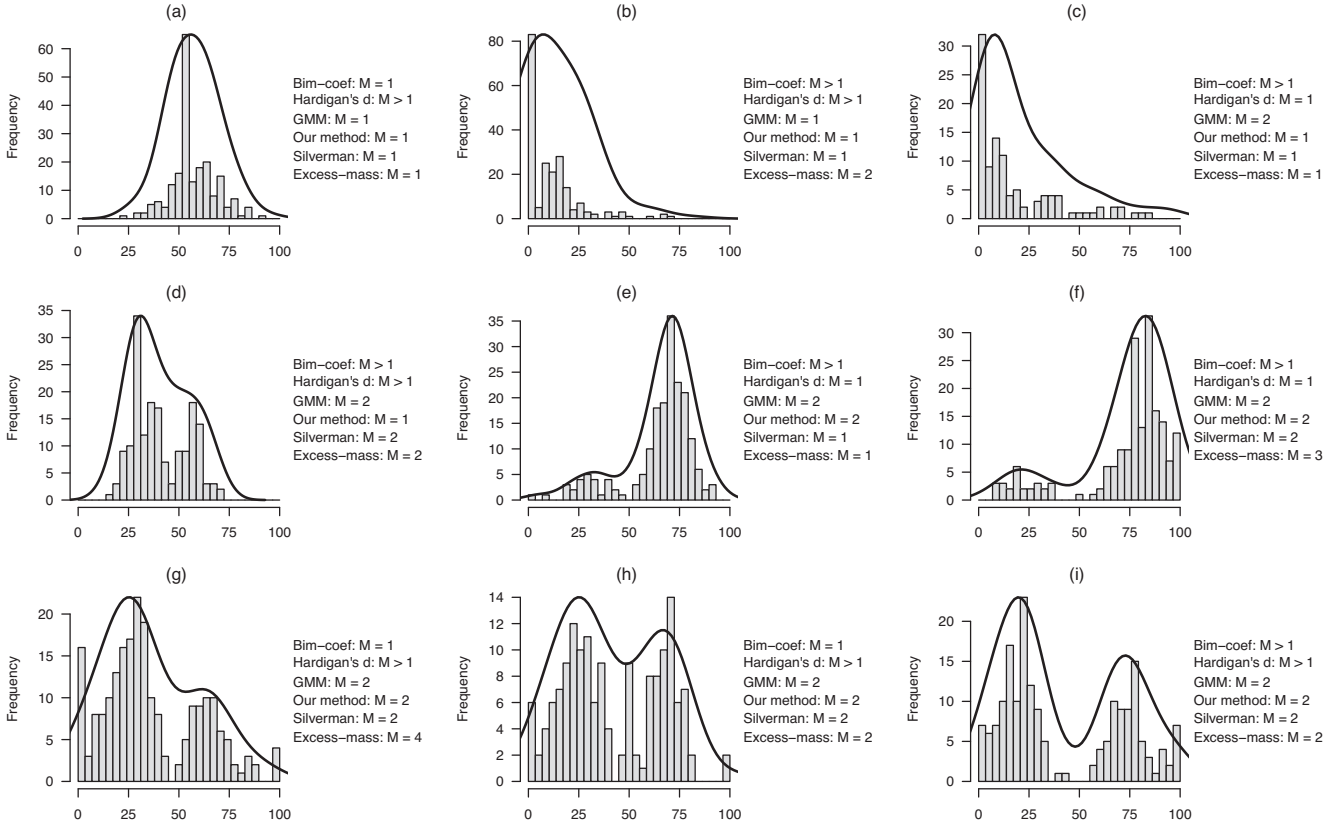
As a comparison, we also applied the bimodality coefficient, Hardigan's dip test, Gaussian mixture modeling, Silverman's method, and an excess mass-based method to the nine examples in Figure A1. The bimodality coefficient identifies multimodality for all cases except (a). This reflects the well-documented weakness of the bimodality coefficient that it cannot distinguish between skewed and multimodal distributions (Pfister et al., 2013). Hardigan's dip test performs very poorly, for example, identifying (a) as multimodal and (f) as unimodal. The Gaussian mixture model (GMM)-based approach works quite well, however, it is more liberal compared to our method. For example, it classifies (c) as bimodal. Silverman's method works very well in all cases and shows the same predictions as our method. The excess mass-based method performs poorly since it both picks up tails as modes or misses modes if the distribution appears bimodal.

Next, we examine the performance of our method in data with a Likert scale instead of a 0–100 slider. Figure A2 displays nine empirical distributions taken from the data of Bringmann et al. (2013) that illustrate the performance of our method. Panel (a) is clearly unimodal and our method detects this correctly. Panels (b) and (c) are clearly unimodally skewed which is again correctly detected. Panel (d) could be seen as skewed unimodal or bimodal with a second mode on the sixth response category and our method estimates a single mode. The remaining distributions are clearly multimodal and our method correctly predicts more than one mode. Similarly to the 0–100 scale, the bimodality coefficient again performs poorly in

(*Appendices continue*)

**Figure A1**

*Compare Mode Estimates of Different Methods with Visual Inspection For 0–100 Scale*




*Note.* Selected empirical examples taken from the data of Rowland and Wenzel (2020) to illustrate the performance of our modality detection method. The black line indicates the density estimate from which we compute the modality.

distinguishing between skewed and multimodal cases. Hardigan's dip test again performs very poorly since it classifies all distributions as multimodal. The Gaussian mixture method performs poorly, since it classifies (e) and (h) as unimodal. Similarly, Silverman's method does not pick up multimodality in (f). The excess mass-based method performs very well for the ordinal scale.

Taking the empirical examples from both scales together, our method performs best since it is the only one that provides mode-estimates that line up with visual inspection for both scales.

### *Appendix A.2.2. Validation via Simulation*

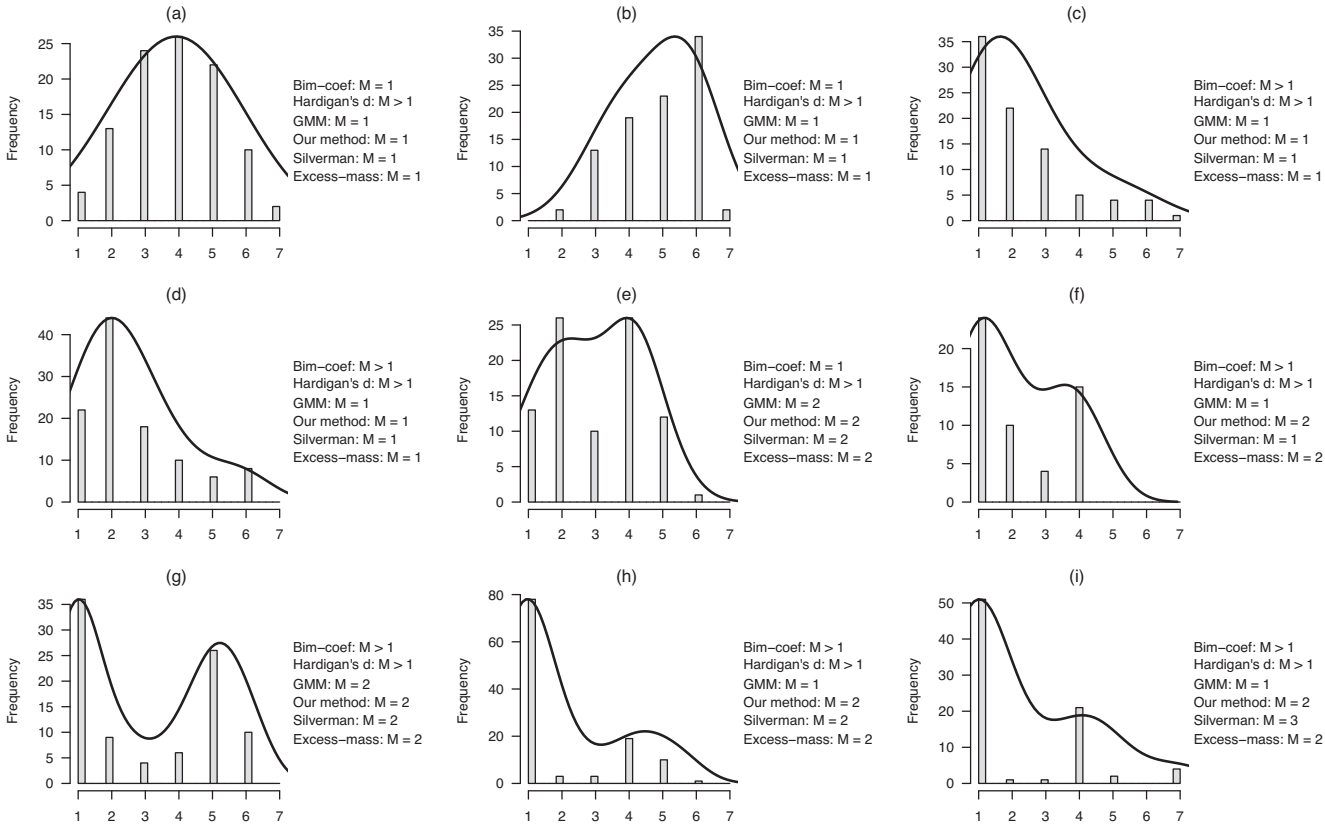
As an additional validation of our method, we used a simulation approach in which we simulated data from symmetric unimodal, skewed unimodal, bimodal, or trimodal distributions (see the top panel in Figure A3). All four distributions are defined as multinomial distributions with $J = 100$ categories, whose $J = 100$ probabilities are defined in the following way. The symmetric unimodal distribution is a Gaussian Mixture Model with $K = 1$ components with mean $\mu = 0$ and $\sigma = 1$. The skewed unimodal distribution is defined by a density based on a power law $x^{-0.4}$. We chose this distribution because it best matches two characteristics of the empirical distributions one would visually classify as unimodal skewed: a large point mass on the lowest category and observations scattered over the rest of the scale, which are best modeled as a fat tail in a unimodal skewed distribution. The bimodal distribution is a $K = 2$ Gaussian mixture model with means ($\mu_1 = 0$, $\mu_2 = 3.5$) and standard deviations $\sigma_1 = \sigma_2 = 1$; and the distribution with three modes consists of $K = 3$ components with means ($\mu_1 = 0$, $\mu_2 = 3.5$, $\mu_3 = 7.5$) and standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = 1$. The mixing probabilities are $1/K$ for both mixtures. In order to map these continuous distributions to the 0–100 scale in the empirical data, we take 100 equally spaced samples from the continuous densities, normalize them, and use them as probabilities in a multinomial distribution. We then sample the data from this multinomial distribution. In the simulation-based validation, we had to exclude Silverman's method and the excess mass-based methods due to the fact that their computational cost was several orders of magnitude higher than the other methods and would have rendered the simulation study unfeasible.

We evaluate the probability that a method correctly classifies a distribution as unimodal versus multimodal (i.e., the method's accuracy). While our density-based method and the GMM-based methods are also able to determine the number of modes, we chose this performance measure in order to compare them against Hardigan's dip test and the bimodality coefficient

(*Appendices continue*)

**Figure A2**

*Compare Mode Estimates of Different Methods with Visual Inspection For 1–7 Scale*




*Note.* Selected empirical examples taken from the data of Bringmann et al. (2013) to illustrate the performance of our modality detection method. The black lines indicate the density estimates from which we compute the modality.

which only distinguishes between unimodality and multimodality. Figure A3 displays the average performance over 1,000 runs as a function of sample size, where we choose variations in sample size that correspond to the range of sample sizes in our data (see Figure 2).

We see that all methods except Hardigan's dip test show high performance in correctly classifying the symmetric unimodal distribution. For the unimodal skewed distributions, Hardigan's dip performs highly, our density-based method performs moderately, and the bimodality coefficient and the GMM perform extremely poorly. For the multimodal distributions, we see that all methods improve in performance with increasing sample size, however, the GMM-based method has the highest performance, followed by our density-based method, Hardigan's dip test, and the bimodality coefficient. These results show that our density-based method works best across the four scenarios.

Figure A4 displays the corresponding results for 7-point Likert-scale data. The symmetric unimodal, skewed unimodal, bimodal, and trimodal distributions are defined by multinomial models with probabilities $\pi_a = \{0.1, 0.2, 0.3, 0.4, 0.3, 0.2, 0.1\}$, $\pi_b = \{0.44, 0.2, 0.13, 0.10, 0.07, 0.04, 0.02\}$, $\pi_c = \{0.1, 0.6, 0.3, 0.1, 0.3, 0.6, 0.1\}$, and $\pi_d = \{0.7, 0.2, 0.3, 0.6, 0.3, 0.2, 0.7\}$. We again report the average accuracy across 1000 runs as a function of sample size.

For the symmetric unimodal distribution, we see that the bimodality coefficient and our density-based method show perfect performance. The GMM performs moderately and Hardigan's dip test performs extremely poorly. For the skewed unimodal distribution, our density-based method performs extremely well, while all other methods perform poorly. For the multimodal distributions, all methods show high performance for a high sample size, however, the methods grow at different rates with sample size. Hardigan's dip test and the GMM-based methods perform best, followed by the bimodality coefficient and our density-based methods. We see that our method performs extremely well in correctly classifying unimodal distributions as unimodal and it shows reasonable sensitivity to detect multimodal distributions. Our method is especially to be desired if one aims to minimize incorrectly classifying unimodal distributions as bimodal, since the method essentially errs only in underestimating the number of modes.

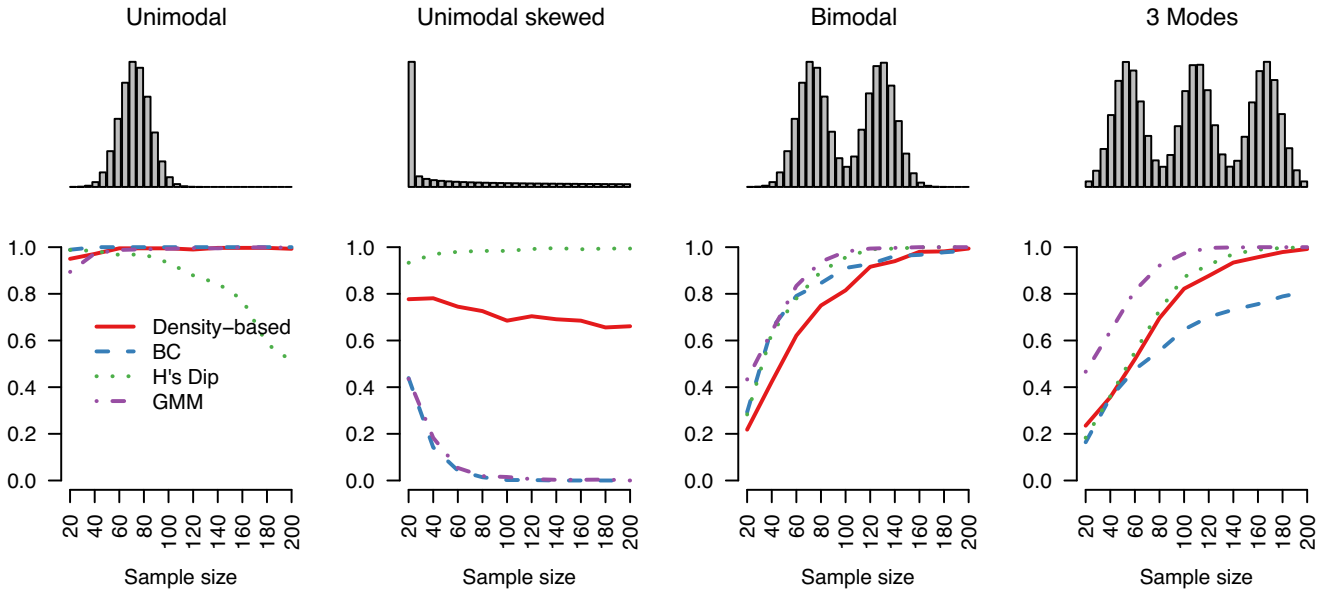
## Appendix A.3. Validation of Skewness Classification

We used our skewness cutoff of 2/3 for the simulation reported in the "reported in the previous section and" section and computed the proportion of correct classifications for symmetric and skewed distributions, both for the 0–100 scale and the Likert scale data. Figure A5

(*Appendices continue*)

**Figure A3**

*Performance Comparison of Different Mode Estimation Methods Using Simulated 0–100 Scale Data*

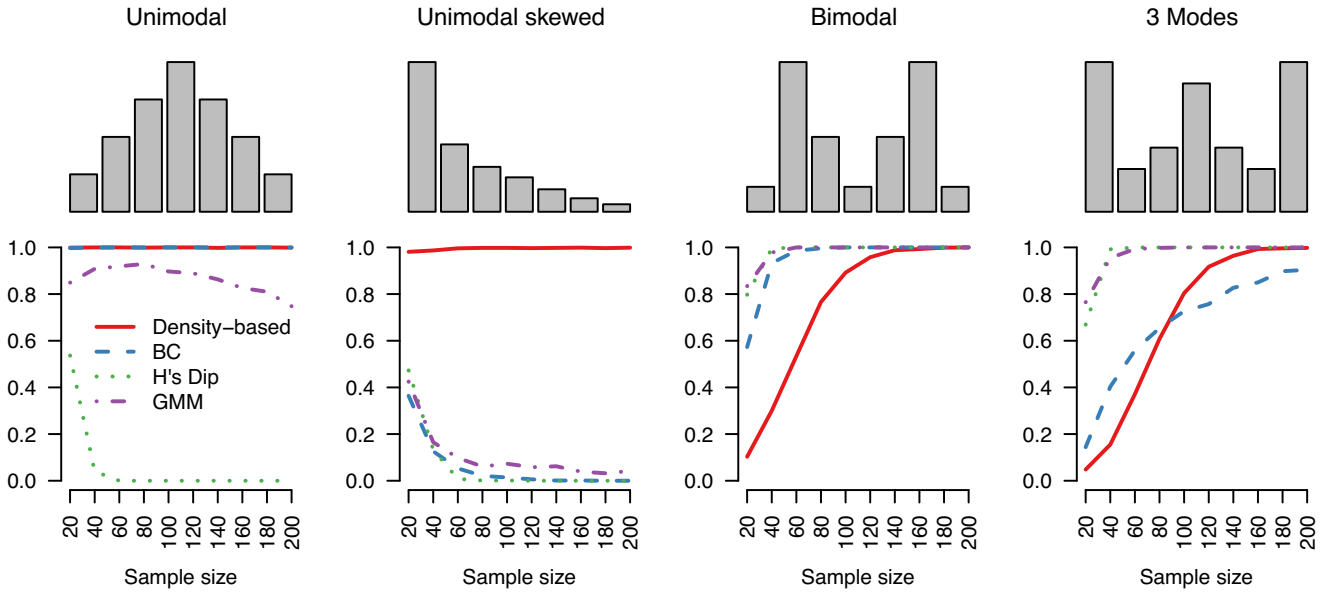



*Note.* The Accuracy of Our Density-Based Method, the Bimodality Coefficient, Hardigan's Dip Test, and the Gaussian Mixture Model (GMM)-Based Density Method for Symmetric Unimodal, Skewed Unimodal, Bimodal, and Trimodal Distributions, as Function of Sample Size. The barplot shows 30 categories for better visibility, but the generated data has 100 categories as in the empirical data analyzed in the main text. See the online article for the color version of this figure.

shows that using the cutoff at 2/3 almost never misclassifies a symmetric distribution as skewed. For the 7-point scale, also symmetric distributions are hardly ever misclassified as symmetric distributions. However, the 0–100 scale skewed distributions in the simulation are often misclassified as symmetric. This shows that the cutoff is conservative, as we claimed in the main text.

**Figure A4**

*Performance Comparison of Different Mode Estimation Methods Using Simulated 1–7 Scale Data*

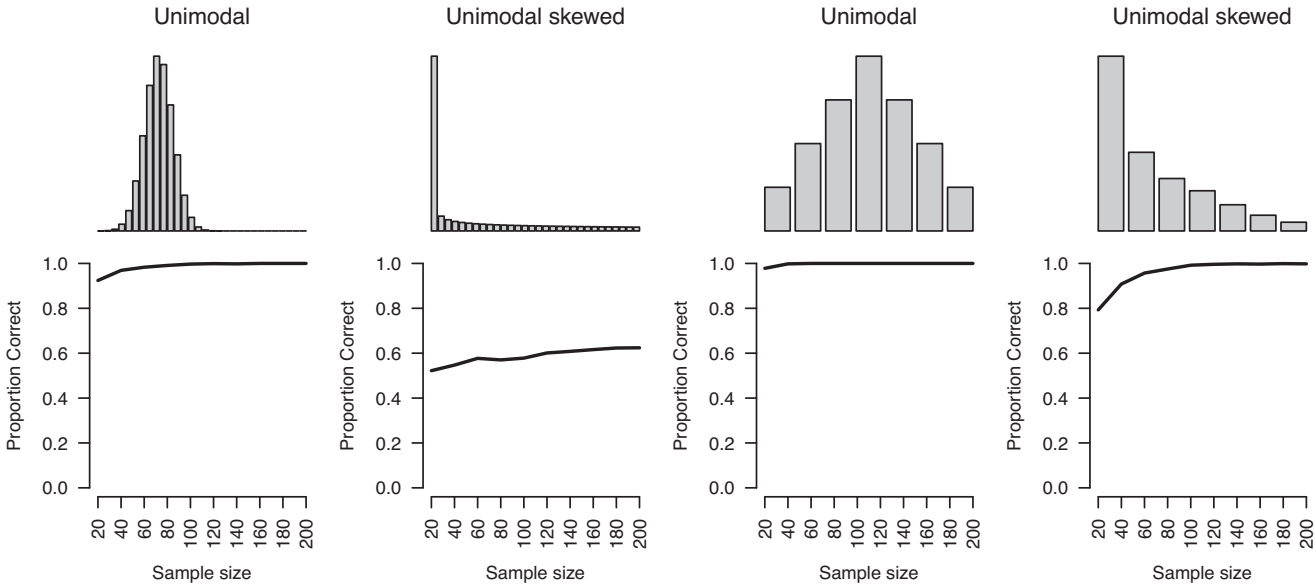



*Note.* The Accuracy of Our Density-Based Method, the Bimodality Coefficient, Hardigan's Dip Test, and the Gaussian Mixture Model (GMM)-Based Density Method for Symmetric Unimodal, Skewed Unimodal, Bimodal, and Trimodal Distributions, as Function of Sample Size for Likert-Scale Data. See the online article for the color version of this figure.

(*Appendices continue*)

**Figure A5**

*Classification Based on Skewness Cutoff of 2/3*




*Note.* Proportion of Correctly Classifying the Symmetric and Skewed Distributions in the Simulation in Section Appendix 2.2 Using a Skewness Cutoff of 1.

## Appendix B

## Data Exclusions Prior Modality Detection

We excluded the responses to items (a) if there was no data, (b) if the standard deviation was below 0.01, or (c) if there were two or fewer unique responses. We excluded these cases because they do not allow a meaningful test for multimodality. Table B1 shows that these criteria only led to exclusions in the datasets of Fried et al. (2020) and Wright et al. (2017) that used a 1–5 Likert scale.

**Table B1**

*Percentage of Excluded Data for Different Exclusion Criteria*

| Data set | No data (%) | $SD < 0.01$ (%) | Range = 1 (%) | Exclusion (%) |
|---|---|---|---|---|
| Rowland and Wenzel (2020) | 0 | 0 | 0 | 0 |
| Bringmann et al. (2016) | 0 | 0 | 0 | 0 |
| Vrijen et al. (2018) | 0 | 0 | 0 | 0 |
| Fisher et al. (2017) | 0 | 0 | 0 | 0 |
| Bringmann et al. (2013) | 0 | 0 | 0 | 0 |
| Fried et al. (2020) | 0 | 0 | 1 | 1 |
| Wright et al. (2017) | 0 | 6 | 11 | 17 |

## Appendix C

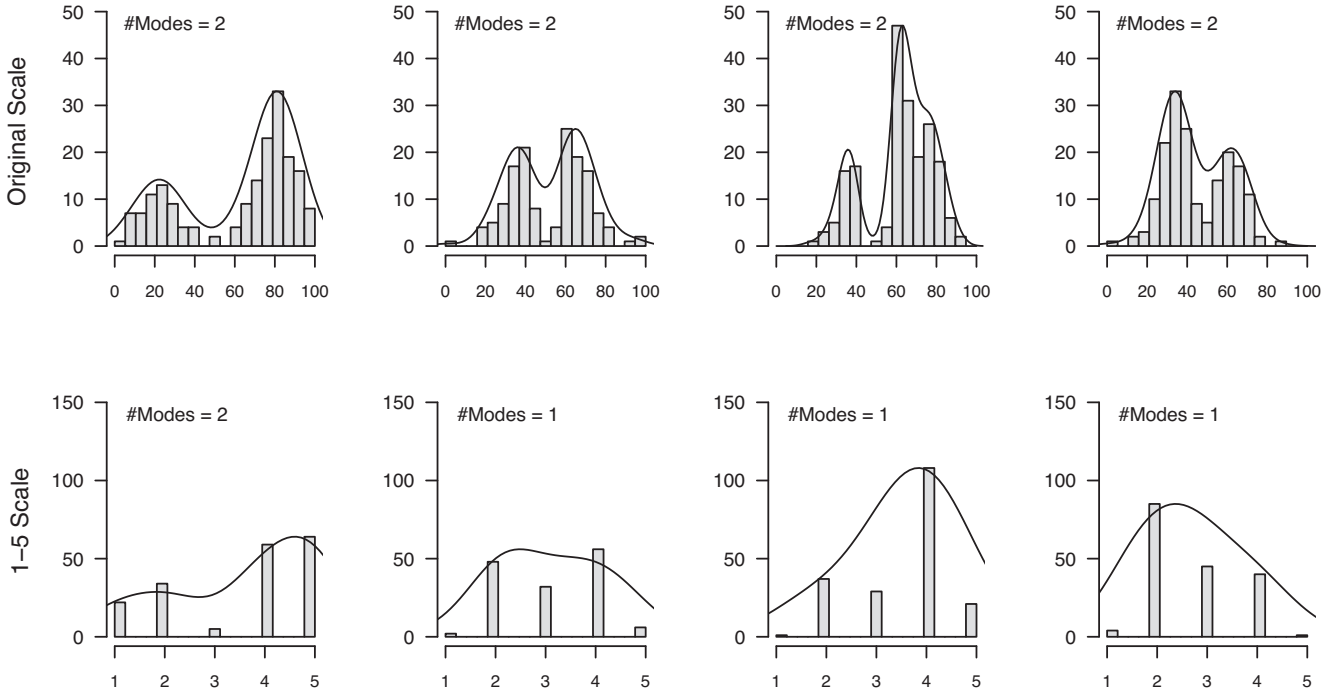
## How Likert Scales Can Mask Multimodality

Figure C1 illustrates how observing only a few ordinal categories may mask the presence of multimodality. In the top row, we display four empirical distributions taken from the data of Rowland and Wenzel (2020), which have been labeled as bimodal in the main analysis. We then threshold the data in five equally spaced intervals between 0 and 100 and label the categories 1, …, 5, which we display in the bottom row. From inspecting the histograms it is already clear that bimodality is largely masked. Only in the example of the first column bimodality is still clearly present on the 1–5 scale.

(*Appendices continue*)

**Figure C1**

*Illustration of How Thresholding can Mask Multimodality*




*Note.* Top Row: Four distributions taken from the dataset of Rowland and Wenzel (2020). Bottom row: the data in the top row thresholded into five equally spaced intervals.

We applied our density-estimation-based modality estimation method to those thresholded Likert-scale data (see black density estimation line) and indeed only find bimodality for the distribution in the first column. This shows that bimodality may be retained if the modes are roughly equally large and are well-separated, but that it is likely masked if this is not the case. Note that this issue is even more problematic for more than two modes. In fact, with a 1–5 scale the maximum number of modes one could theoretically observe is 3.

We would like to note that the thresholding using equally spaced intervals which we applied here does not necessarily reflect how individuals would choose ordinal answer categories based on their score on a hypothetical latent continuous scale. If an individual is aware of the presence of qualitatively different intensities, they might intuitively choose a mapping to the ordinal scale that allows them to express the discrete nature of their experience. Still, the argument holds that ordinal scales with few categories have the tendency to mask multimodality which is an argument for using the 0–100 scale.

## Appendix D

## Analysis of Multimodality

As described in the main text, we sequentially built more complex models and assessed their comparative performance using the BIC. All analyses were conducted in Julia using the *MixedModels.jl package*.[2] Table D1 provides an overview of the models.

We first fit a null model consisting of random-intercepts on the person and study level. The next step of our analysis is to examine if any available predictor variables explain the variation in multimodality across items, individuals, and studies. On the item level (level 1) we have a single predictor—the valence (negative or positive and neutral) of the item. Adding valence as a level-1 (i.e., item-level) predictor (Model 2), we see a significant improvement in model fit only if we additionally add random slopes on the item and person level (Model 3), encoding the assumption that the effect of valence differs across items and persons. Next we examine whether any level-2 predictors can explain the variation in multimodality we observe across individuals and studies. Although each of the datasets includes a variety of different between-person characteristics of their subjects, such as

[2] We used Julia because running the models in R would have either taken a substantially longer time or led to non-convergence.

(*Appendices continue*)

**Table D1**
*Model Selection Results for the Modality Data*

| | Model | AIC | BIC | $\chi^2\Delta(df)$ |
|---|---|---|---|---|
| | Null model | | | |
| 1 | Null model RI Levels 1 and 2 | 5,001 | 5,023 | |
| | Level-1 predictors | | | |
| 2 | + valence | 5,000 | 5,029 | 2.70(1), $p = .10$ |
| 3 | + valence (RS Levels 1 and 2) | 4,707 | 4,764 | 301.84(4), $p < .001$ |
| | Level-2 predictors | | | |
| 4 | + $n_{obs}$ | 4,702 | 4,767 | 6.46(1), $p = .01$ |
| 5 | + $n_{obs}$ × valence | 4,678 | 4,750 | 25.88(1), $p < .001$ |
| | Level-3 predictors | | | |
| 6 | + scale + scale × valence − RS valence — study | 4,674 | 4,745 | NA, NA |
| 7 | + (MPD + retro + pop) × valence | 4,670 | 4,784 | 15.44(4), $p = .02$ |

*Note.* AIC = Akaike information criterion; BIC = Bayesian information criterion; RI = random intercepts; RS = random slopes; Valence = valence of items, coded 0 for negative items, 1 for positive; $n_{obs}$ = number of observations for that individual, centered around the average number of observations across all people in all studies; retro = phrasing of items, coded 0 if participants asked about current emotion and 1 if asked about emotion since the last measurement occasion; scale = dummy coded as 0 if a Likert scale of 5 or 7 values, 1 if a slider of 100; MPD = mean number of measurements per day in the study design, centered around the mean across studies; pop = dummy coded if the study consisted of a clinical population or student population. Each model adds new predictors to the previous model. The columns denoted $\chi^2\Delta(df)$ show the $\chi^2$ difference test statistics, degrees of freedom and *p*-values associated with testing the model on any given row with the model specified in the directly preceding row. When examining models with Level-3 predictors, we add one predictor at a time and chose the model with the greatest improvement according to the BIC. Model 7 represents the model with all Level 3 predictors and all interactions with valence.

neuroticism, there is no such variable that is measured in all studies. As such, the only between person characteristic we can use in this analysis is the number of observations or measurements of each person. We denote this as $n_{obs}$, and for interpretability we center this variable around the mean across persons of 100.86. As we can see from Table D1, adding the main effect of $n_{obs}$ and its interaction with valence improves the model fit (Model 5). As a final step, we investigate the degree to which study-level predictors could explain variance in multimodality. We find that adding the predictor scale (indicating whether it is a Likert scale of 5 or 7 values or a slider) and its interaction with valence improves the fit (both in terms of AIC and BIC) if we also remove the random slope of valence across studies (Model 6; otherwise the fit does not improve).[3] Lastly, the model improves when adding the mean number of measurements per day, the phrasing of the items (asking about the current emotion or since the last measurement occasion), and the study population (clinical or student), constituting our final model (Model 7).

## Appendix E

## Skewness Analysis Details

Analyses were conducted using the lme4 package in R (Bates et al., 2015). Parameter estimates were obtained using REML, while model comparisons are computed based on maximum likelihood model fit. We first fit a null model consisting of random intercepts on the person and study level. The next step of our analysis is to examine if any available predictor variables explain the variation in skewness across items, individuals, and studies. On the item level (Level 1) we have a single predictor, that is, the valence (negative or positive/neutral) of the item. Based on Figure 5, we would expect that the valence would be a strong predictor of skewness, and this is born out by our analysis: adding valence as a predictor, and allowing the effect of valence to differ across participants and studies (Model 2 in Table E1), yields an improved fit.

Next we examine whether any level-2 predictors can explain the variation in skewness we observe across individuals and studies. As we can see from Table E1, adding the main effect of $n_{obs}$ and its interaction with valence improves the model fit (Model 5). This is the final model reported in the main text.

The final model exhibits random intercepts and random slopes, with a strong negative correlation between them, at both levels: The higher the skewness of negative items, the larger the difference between the skew of negative and positive items. We can understand this by considering the mean skewness of positive and negative items for each person, and how the relationship between these quantities differs across studies, as shown in Figure E1. We can see that there is a positive relationship between the average skewness of positive and negatively valenced items

[3] Because the models are not nested anymore after removing the random slope of valence across studies, we cannot calculate *p*-values. This is indicated as NA in the Table D1.

(*Appendices continue*)

**Table E1**
*Model Selection Results for the Skewness Data*

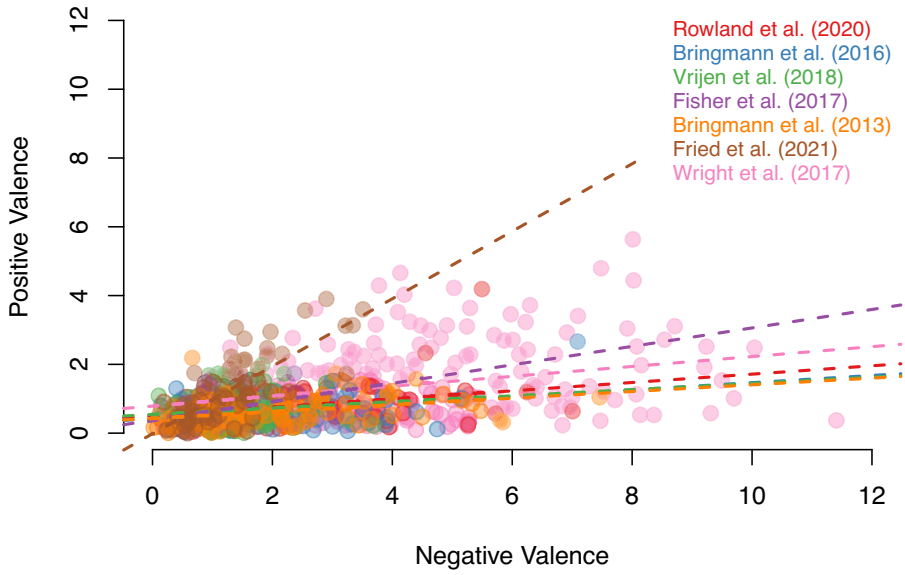
| | Model | AIC | BIC | $\chi^2\Delta(df)$ |
|---|---|---|---|---|
| | Null model | | | |
| 1 | RI levels 1 and 2 | 3,3096 | 3,3124 | |
| | Level-1 predictors | | | |
| 2 | + valence | 30,826 | 30,861 | 2,272(1), $p < .001$ |
| 3 | + valence (RS Levels 1 and 2) | 29,208 | 29,271 | 1,626(4), $p < .001$ |
| | Level-2 predictors | | | |
| 4 | $+ n_{obs}$ | 2,9198 | 2,9268 | 11.9(1), $p < .001$ |
| 5 | $+ n_{obs} \times$ valence | 2,9151 | 29,228 | 48.5(1), $p < .001$ |
| | Level-3 predictors | | | |
| 6 | + retro + retro × valence | 2,9143 | 29,233 | 12.5(2), $p = .002$ |
| 7 | + scale + scale × valence + MPD + MPD × valence + pop + pop × valence | 29,144 | 29,270 | 9.2(6), $p = .16$ |

*Note.* AIC = Akaike information criterion; BIC = Bayesian information criterion; RI = random intercepts; RS = random slopes; MPD,= mean number of measurements per day in the study design, centered around the mean across studies. Model 7 represents the model with all Level 3 predictors and all interactions with valence.

across people and that this relationship differs in strength across studies. Furthermore, the range of skewness values for negatively valenced items is considerably larger than for positively valenced items. This means that as the skewness of negative items gets larger, the skewness of positive items changes less strongly: The covariance between the random effects is negative since the larger the skew of the negative items, the greater the difference between the skew of negative and positive items. The effect of $n_{obs}$ can be interpreted as showing that the effect of increasing the time series length is much greater for negatively valenced items than for positive ones. This implies that on average positively valenced emotions have a less skewed distribution and that our ability to detect skewness of negatively valenced items may be more strongly affected by the time series length.

As a final step, we investigated the degree to which study-level predictors could explain variance in skewness. Adding any combination of Level-3 predictors decreases fit according to the BIC, but with inconsistent changes in model fit according to the AIC and $\chi^2$ difference test. The main results of a step-wise model search procedure are shown in Table E1, where we see that adding the retrospective phrasing variable *retro* (whether the items queried emotion since the last measurement occasion vs. current emotion) improved model fit according to the AIC and $\chi^2$ difference test, but decreased model fit according to the BIC. We do not report the parameters of the model here, since qualitatively the effects of valence and time-series length remain the same, while the effects of retrospective phrasing are nonsignificant. Adding any other combination of Level-3 predictors decreased model fit according to the AIC.

**Figure E1**
*Mean Skewness of Positive and Negatively Valenced Items per Person, Across Studies*




*Note.* See the online article for the color version of this figure.

(*Appendices continue*)

### Appendix E.1. Study-Specific Analysis

In Table E2, we show the parameter estimates of the final models for Bringmann et al. (2013), Bringmann et al. (2016), and Wright et al. (2017). Note that for Bringmann et al. (2016) the model with the main effect and interaction of neuroticism did not consistently improve model fit in comparison to the model with only an intercept and valence as a predictor: The model with neurotic had lower AIC, but larger BIC and a nonsignificant $\chi^2$ difference test ($\chi^2(2) = 4.913$, $p = .086$).

**Table E2**

*Summary of Subject-Specific Analysis for Bringmann et al. (2013), Bringmann et al. (2016), and Wright et al. (2017)*

| Height | Bringmann et al. (2013) | Bringmann et al. (2016) | Wendt et al. (2020) |
|---|---|---|---|
| Intercept (*SE*) | 1.93*** (0.09) | 1.81*** (0.11) | 4.13*** (0.10) |
| Valence (*SE*) | −1.31*** (0.09) | −1.13*** (0.13) | −2.70*** (0.06) |
| Neurotic (*SE*) | −0.32*** (0.09) | −0.24* (0.12) | −0.88*** (0.10) |
| Valence × Neurotic (*SE*) | 0.29** (0.09) | 0.23 (0.14) | 0.89*** (0.06) |

*Note.* In all studies, neuroticism was standardized around the grand mean.
*$p < .05$. **$p < .01$. ***$p < .001$.

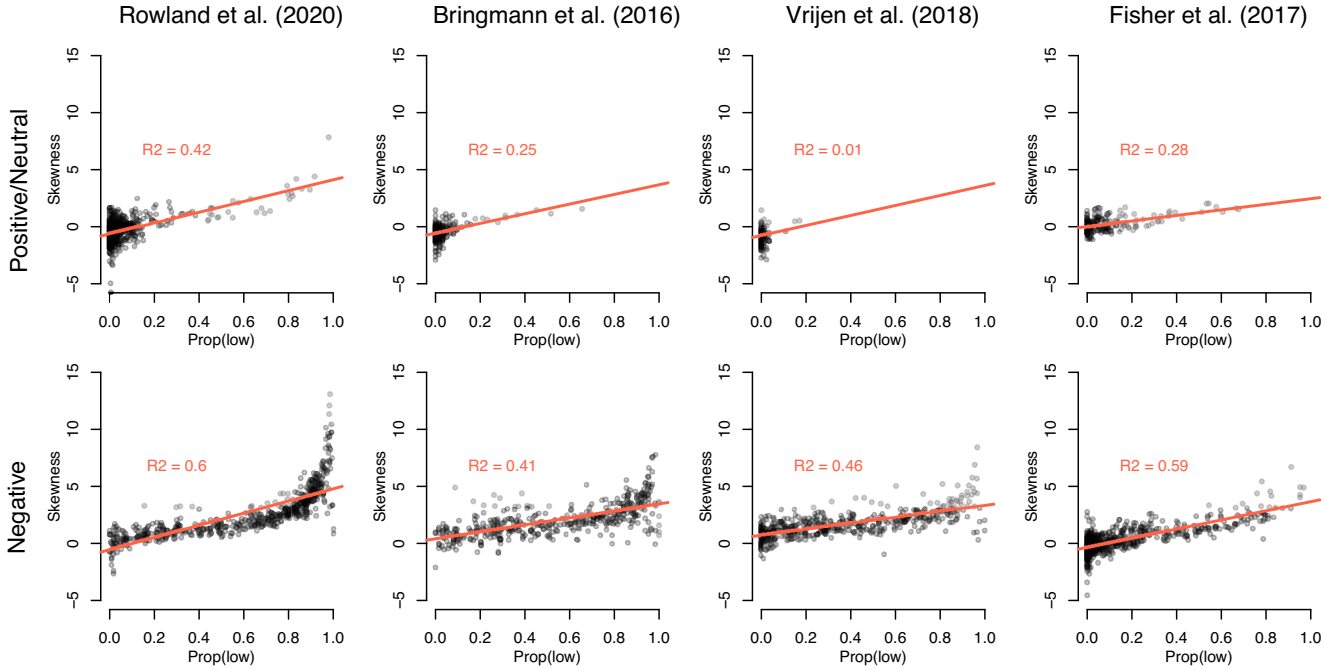
## Appendix F

## Relation Between Skewness and Distribution Across Scale

In the main text, we investigate the relationship between skewness and item, participant, and study characteristics. In this appendix, we examine the degree to which the skewness of an item correlates with the tendency for responses to cluster around the bottom of the rating

**Figure F1**

*Relation Between Skew and Frequency of Lowest 10% Quantile for Positive and Negative Emotions in Studies With a 0–100 Scale*

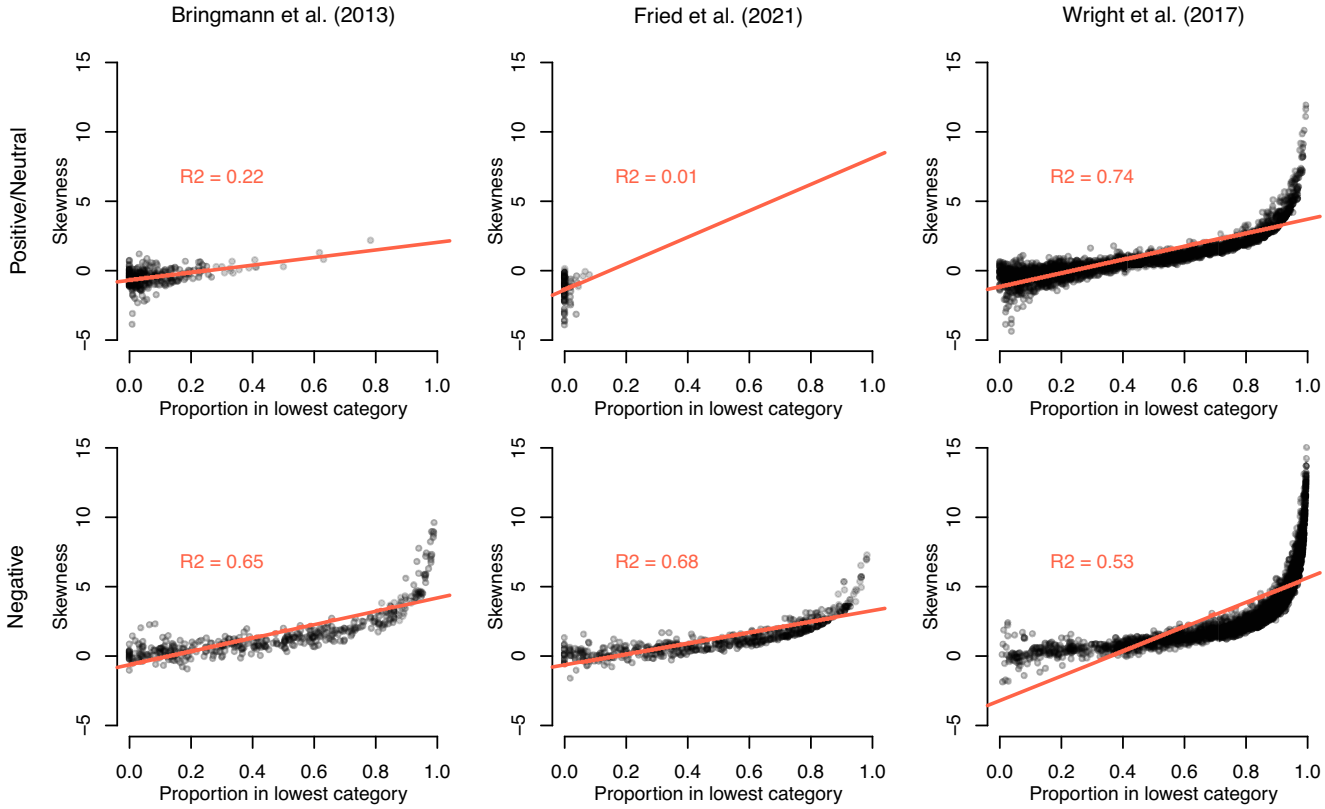



*Note.* Proportion of observations in the lowest 10% quantile (*x*-axis) and skewness (*y*-axis) for positive/neutral (top) and negative (bottom) items across four studies. See the online article for the color version of this figure.

(*Appendices continue*)

**Figure F2**

*Relation Between Skew and Frequency of Lowest Category for Positive and Negative Emotions in Studies With Likert Scales*




*Note.* Proportion of observations in the lowest category (*x*-axis) and skewness (*y*-axis) for positive/neutral (top) and negative (bottom) items across three studies. See the online article for the color version of this figure.

scale. To operationalize this, we calculate the proportion of measurements (a) in the lower 10% quantile for 0–100 scales or (b) which endorse the lowest ordinal category for Likert scale items. Figure F1 shows the results for the four studies with a 0–100 scale, separately for positive, neutral, and negative emotions. Figure F2 shows the results for the three studies with Likert scales, separately for positive and neutral and negative emotions. A similar result emerges when using the mean instead of the proportion measure (but showing a negative rather than positive relation with skewness).

## Appendix G

## Assessing Model Fit of VAR Model via Simulation

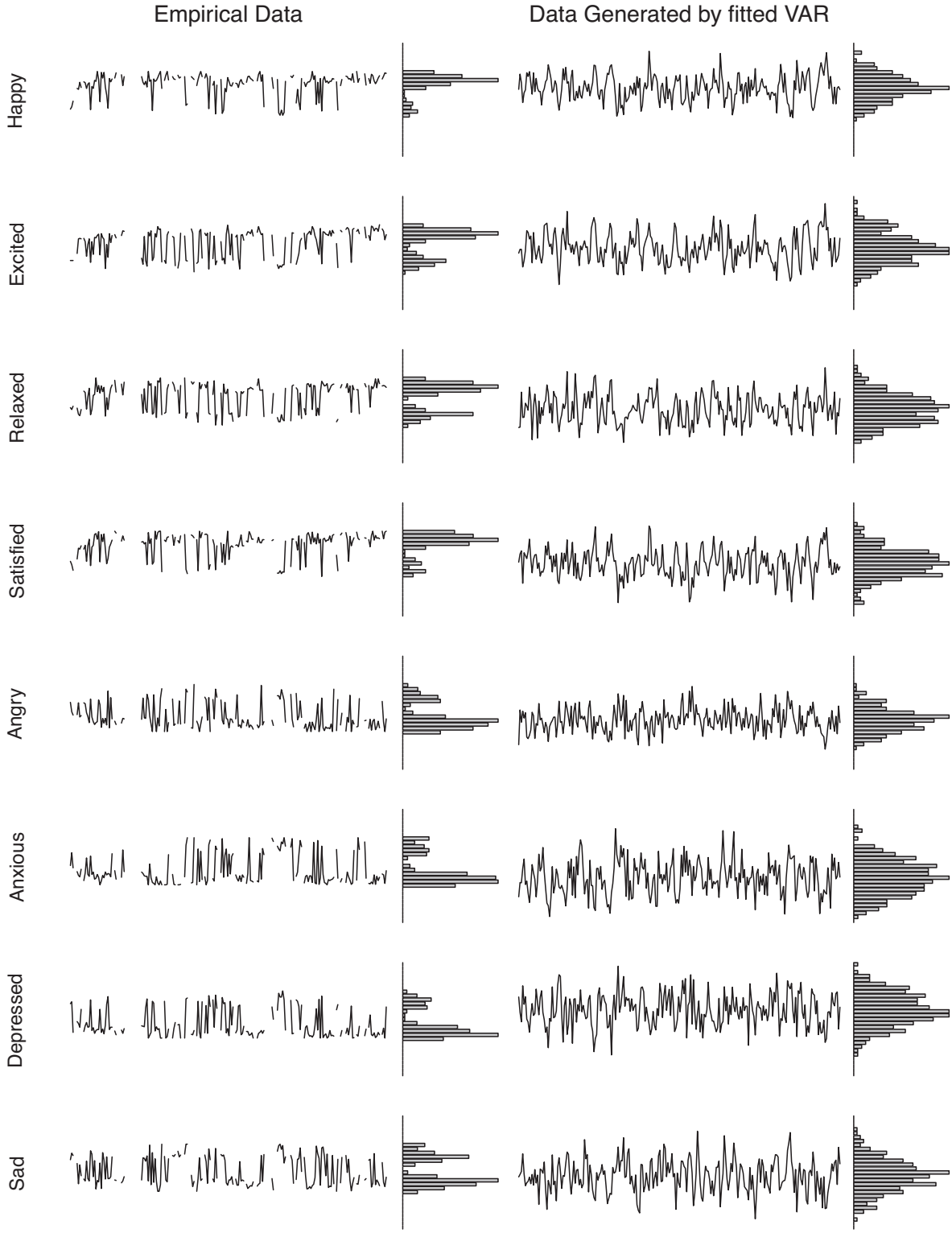
Here we fit a VAR model to the eight emotion variables of the first individual in the dataset of Rowland and Wenzel (2020) and then generate data from the VAR model using the data at the first time point of the empirical data as initial values. Figure G1 shows the empirical data (black lines) and the data generated from the fitted VAR model for each of the eight emotions.

We see that the empirical time series looks very different from the one generated by the fitted VAR model. The empirical time series seems to jump between different equilibria, while the simulated data varies around a single equilibrium. Note that there is more data in the simulated time series, because there are no missing time points. Note that the initial time points in the plot sometimes appear not to

(*Appendices continue*)

**Figure G1**

*Left Panel: Empirical Data of Eight Emotion Variables of the First Subject in the Data of Rowland and Wenzel (2020). Right Panel: Data Generated From a Vector Autoregressive (VAR) Model Estimated on the Empirical Data in the Left Panel*



be the same in the empirical and the simulated data. This is the case when the second time point is missing and consequently no line is plotted between the first and the second time point.

Received May 16, 2022
Revision received October 18, 2022
Accepted December 19, 2022 ■