UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Ecker, S., Ottenstein, C., Vollbracht, D., & Lischetzke, T. (2026). Does the Procedure Matter? Applying a Multiverse Analysis Approach to Negative Emotion Differentiation. *Collabra: Psychology*, *12*(1). https://doi.org/10.1525/collabra.159939

Methodology and Research Practice

Does the Procedure Matter? Applying a Multiverse Analysis Approach to Negative Emotion Differentiation

Sabrina Ecker[1][a], Charlotte Ottenstein[2], Dominik Vollbracht[1], Tanja Lischetzke[1]

[1] Department of Psychology, RPTU University Kaiserslautern-Landau, Landau, Germany, [2] Center for Methodologies, Diagnostics and Evaluation, RPTU University Kaiserslautern-Landau, Landau, Germany

Keywords: Multiverse Analysis, Negative Emotion Differentiation, Researcher Degrees of Freedom, Robustness, Research Transparency

https://doi.org/10.1525/collabra.159939

Collabra: Psychology

Vol. 12, Issue 1, 2026

Negative emotion differentiation (NED) is frequently assessed using momentary emotion ratings in ambulatory assessment studies. However, researchers differ in how they preprocess data in studies on NED, and this variation may affect empirical findings. The present research scrutinized whether decisions in NED data processing affect the robustness of NED's relationships with adaptive outcomes using multiverse analysis—that is, conducting the analysis of interest across all reasonable combinations of methodological decisions. We included decisions on the compliance threshold for exclusion, the inclusion of occasions, and how the NED index was calculated. The analyses of interest were the bivariate between-person correlation between NED and depression and, using multilevel analysis, the buffering effect of NED on daily stress reactivity. For both analyses of interest, separate multiverse analyses were conducted on three ambulatory assessment data sets (163 to 406 participants, 11,876 to 21,552 occasions) collected in Germany between 2020 and 2024. The results indicated that the bivariate between-person correlation between NED and depression was not robust in any of the data sets. The buffering effect of NED on daily stress reactivity was robust in one data set. The impact of specific decisions varied across data sets and analyses of interest. With the exception of one decision, each decision was identified as the most influential in at least one multiverse analysis. However, there was no systematic overall pattern in how the decisions affected the outcomes. The results of the multiverse analyses highlight the importance of research transparency and careful data preparation, as methodological decisions may affect empirical findings on NED. Whether the results generalize to other analyses of interest and to more diverse or non-German samples remains to be determined.

Emotion differentiation (also labeled emotional granularity) describes the extent to which individuals distinguish between like-valence emotional states in a fine-grained manner (Feldman Barrett et al., 2001; Kashdan et al., 2015). Individuals with high emotion differentiation tend to use discrete emotions terms such as "angry" or "sad" in a situation-specific way. In contrast, individuals with low emotion differentiation tend to report similar levels of those emotions across situations, indicating that they use emotion terms in a less context-sensitive and more interchangeable manner (Erbas et al., 2019; Lischetzke et al., 2021). In particular, distinguishing between negative emotions (*negative emotion differentiation*; NED) is considered beneficial for psychological well-being (Kashdan et al., 2015), and studies have provided evidence for small to moderate associations between higher NED and adaptive outcomes in non-clinical samples (e.g., lower levels of depressive symptoms, Erbas et al., 2014, 2019; Starr et al., 2017, 2020; Willroth et al., 2020; reduced stress reactivity, Lischetzke et al., 2021; Nook et al., 2021; Starr et al., 2017, 2020; higher behavioral adaptation, O'Toole et al., 2020; lower enactment of maladaptive behaviors, Seah & Coifman, 2022).

NED is typically measured through intensive longitudinal designs (ambulatory assessment [AA]) in which participants repeatedly rate their momentary emotional experience using a fixed set of emotion items. Individuals with higher NED tend to use emotion terms in a more discerning manner, resulting in more independent ratings of different emotions over time and, consequently, lower covariation among emotion ratings. In contrast, individuals with

a Correspondence concerning this article should be addressed to Sabrina Ecker, Department of Psychology, RPTU University Kaiserslautern-Landau, Fortstraße 7, 76829 Landau, Germany. Email: sabrina.ecker@rptu.de

lower NED tend to rate their emotional experiences more uniformly such that changes in one emotion co-occur with changes in others (e.g., feeling anxious and depressed), resulting in higher covariation among emotion ratings (Feldman Barrett et al., 2001; Schreuder et al., 2022). This covariation is typically quantified separately for each participant by calculating the intraclass correlation (ICC) between the ratings of momentary emotions across measurement occasions (Thompson, Springstein, et al., 2021). Originally developed to assess rater reliability, where *k* raters rate *n* targets (Shrout & Fleiss, 1979), the ICC has been adopted for NED research, in which each participant provides *k* emotion ratings across *n* measurement occasions (Ottenstein & Lischetzke, 2020). Higher ICCs reflect greater covariation of emotion ratings across time and, thus, lower NED, whereas lower ICCs indicate more distinct emotion ratings across time and higher NED. Accordingly, ICCs are inverted so that higher values represent greater NED (Thompson, Springstein, et al., 2021).

Because emotion items are typically selected based on theoretical considerations rather than sampled randomly from a broader universe of items, NED research commonly relies on a two-way mixed-effects model (ICC [3]), which treats the included emotion items as fixed and does not aim to generalize beyond them (Ottenstein & Lischetzke, 2020). Although the average inter-item correlation represents an alternative index of NED, this approach is less commonly used (Thompson, Springstein, et al., 2021). Accordingly, the present research focuses exclusively on ICC-based indices.

As highlighted in the review by Thompson, Springstein, and Boden (2021), researchers seeking to quantify individual differences in NED must make several methodological decisions, such as the specific calculation method of the ICC, and whether and how to apply inclusion criteria. Empirical studies differ in the way these issues have been addressed, raising the question of whether and how these methodological decisions affect empirical findings on NED. The aim of the present research was to scrutinize the robustness of the relationship between NED and adaptive outcomes (lower depression[1], reduced momentary stress reactivity) by applying a *multiverse analysis* that considers all reasonable data processing decisions (Steegen et al., 2016). To this end, we analyzed data from three AA studies that measured depression and assessed participants' emotional states and stress reactivity in everyday life. In the remainder of the introduction, we will first provide the theoretical and empirical background for our analyses of interest regarding the adaptive value of NED. Second, we will elaborate on the use of multiverse analysis, and third, we will describe the identified reasonable methodological decisions in NED research that, in combination, form the multiverse of data sets to be analyzed.

The Present Analyses of Interest on NED's Adaptive Value

Negative emotions provide critical information about a situation, potential causes of distressing feelings, and possible behavioral responses (Schwarz, 2012). Theory and empirical research suggest that discerning one's negative emotions in a more differentiated manner facilitates effective emotion regulation (Feldman Barrett et al., 2001; Kalokerinos et al., 2019; Kashdan et al., 2015). Among the adaptive outcomes that high NED is believed to confer through this process are lower depression (Eckland et al., 2022; Starr et al., 2017, 2020; Willroth et al., 2020) and reduced reactivity to daily stress (Starr et al., 2017, 2020). Meta-analyses on the association between NED and indices of behavioral adaptation (O'Toole et al., 2020; Seah & Coifman, 2022) support the idea that experiencing negative emotions, such as embarrassment or sadness, in a more granular way helps individuals resort less to maladaptive regulation strategies, such as aggression or binge drinking (Kashdan et al., 2015). Thus, higher NED should promote daily well-being and, over time, prevent or reduce depressive symptoms. In situations perceived as overwhelming or taxing (i.e., stress exposure), individuals with high NED should be better able to identify the specific emotions triggered by stressors, such as feeling "anxious" versus "irritated" or "disappointed." This detailed emotional awareness should facilitate a more targeted and effective coping response (Starr et al., 2020). Consequently, those with higher NED should be less likely to experience strong or prolonged stress responses (ultimately supporting better mental health outcomes, such as lower depressive symptoms).

Several studies of non-clinical samples have tested whether higher NED is related to lower depressive symptoms (Eckland et al., 2022; Erbas et al., 2014, 2019; Lennarz et al., 2018; Lischetzke et al., 2021; Liu et al., 2020; Matt et al., 2016; Starr et al., 2017, 2020; Williams & Uliaszek, 2022; Willroth et al., 2020). Most findings supported a small to moderate negative correlation between NED and depression, with a few exceptions (Eckland et al., 2022, Sample 2; Lennarz et al., 2018; Lischetzke et al., 2021; Liu et al., 2020, depression at follow-up; Matt et al., 2016). This association held when controlling for mean emotion intensity levels (Starr et al., 2020; Willroth et al., 2020). Thus, in our multiverse analyses, we examined the bivariate relationship between NED and depression as the first analysis of interest, anticipating a small to moderate correlation.

The association between NED and momentary or daily stress reactivity has been tested in five AA studies (Lischetzke et al., 2021; Nook et al., 2021; Starr et al., 2017, Study 1 and Study 2; Starr et al., 2020) using a multilevel approach to operationalize individual differences in reactivity to momentary/daily stress. In the studies by Starr and

1 Throughout the manuscript, we use "depression" to refer to a continuous symptom dimension rather than a categorical diagnosis of major depressive disorder.

colleagues, NED attenuated the positive within-person relation between momentary negative experiences and momentary depressed mood in a sample of veterans (Starr et al., 2017, Study 2) and the positive within-person relation between daily hassles and daily depressive symptoms in a community sample of adolescents (Starr et al., 2020), but not in undergraduates (Starr et al., 2017, Study 1). The results of the significant cross-level interactions remained unchanged when controlling for mean emotion intensity. Nook et al. (2021) found that NED (measured via responses to standardized stimuli) buffered the positive within-person relationship between momentary stress and momentary depressed, but not anxious, affect in a sample of female adolescents, even when controlling for the mean negative affect intensity across stimuli. In a community sample (Lischetzke et al., 2021), individuals with higher NED showed a weaker within-person link between higher daily stress and reduced evening calmness than those with lower NED, and again this cross-level interaction between NED and stress held when controlling for mean emotion intensity. In our multiverse analyses, as our second analysis of interest, we examined the relation between NED and individual differences in momentary stress reactivity using a multilevel model (i.e., we tested the cross-level interaction between NED and momentary stress on calm vs. tense mood). We expected that individuals with higher NED would show a less negative within-person relationship between momentary stress and momentary calmness.

The Multiverse Analysis Approach and Its Rationale

Even when researchers are given an identical data set and research question, differences in how they prepare and analyze the data can produce different results (Silberzahn et al., 2018). This flexibility, known as researcher degrees of freedom, can introduce the risk of inflated α-error and biased conclusions, as researchers might—intentionally or unintentionally—make choices that yield more favorable results (Simmons et al., 2011). To mitigate this, researchers may commit to a single, well-considered data processing approach, often through preregistration. However, even with preregistration, the resulting data set represents only one version ("universe") of many possible within a "multiverse", and results may vary across this multiverse (Steegen et al., 2016). To assess the robustness of empirical findings across methodological decisions, *multiverse analysis* has been proposed as a valuable tool. It involves identifying all methodological decisions and their reasonable options, generating the valid combinations of these (while excluding inconsistent or redundant ones), and "performing the analysis of interest across the whole set of data sets that arise from different reasonable choices for data processing" (Steegen et al., 2016, p. 703). Results are then reported for all universes and summarized using, for example, an outcome curve that plots all universes on the x-axis and their outcome values on the y-axis (Hall et al., 2022) to map the distribution of estimates and identify influential decisions (Simonsohn et al., 2020). Stable results across the multiverse are considered robust (Hall et al., 2022).

Methodological Decisions in NED Research

In NED research, several methodological decisions have to be made that pertain either to the treatment of intensive longitudinal data or to specific choices in the calculation of the ICC. The decisions were identified by reviewing the NED literature and noting those that were frequently described yet handled differently across studies. Most of these have been previously outlined in a literature review by Thompson, Springstein, and Boden (2021). Below, we describe each decision and its options (for an overview, see Table 1). Importantly, we include in the multiverse only those decisions that are central to the estimation of NED itself—that is, decisions that directly influence the derived NED indices. Other methodological choices that researchers must make—such as selecting an appropriate modeling approach for their research question—are beyond the scope of the present multiverse analysis.[2]

Compliance Threshold (Inclusion of Participants)

In intensive longitudinal research on NED, it is argued that a sufficient number of measurement occasions is essential for a reliable NED index (e.g., Erbas et al., 2018; Willroth et al., 2020), but studies differ in the number of occasions used to assess NED (Thompson, Springstein, et al., 2021). Consequently, participants are often excluded due to low compliance (i.e., the number of completed prompts relative to all scheduled prompts; Weermeijer et al., 2022), but the minimum compliance threshold is somewhat arbitrary. Previous studies have varied from no applied threshold (with individuals' actual compliance rates between 20% and 100%; Liu et al., 2020) to a 60% threshold (Erbas et al., 2018). A recent multiverse analysis examined compliance thresholds from 0% to 50% (Weermeijer et al., 2022). However, this study did not consider more conservative thresholds or focus on NED. To cover a broad range of compliance thresholds while ensuring the interpretation of the multiverse analysis remained manageable, we applied four compliance thresholds: 12.5%, 25%, 50%, and 75%. We included

2 In our preregistration, we included a decision to vary the emotion item set used to calculate NED based on the empirical mean intensity of emotions. This approach was motivated by prior work highlighting item selection as a critical decision in NED research (Thompson, Springstein, et al., 2021). However, when incorporating three data sets in the present research, this method resulted in inconsistent classifications of emotion terms (e.g., fear was classified as a high-intensity emotion in Study 1, but as a low-intensity emotion in Study 2). In the absence of German affective norms applicable to all emotion terms used across the three studies, we ultimately omitted this decision from the final analyses.

Table 1. Overview of the Decisions and their Corresponding Options in the Multiverse Analyses

Option Nr.	Option
Decision 1: Compliance threshold (inclusion of participants)	
1	At least 12.5% of occasions completed
2	At least 25% of occasions completed
3	At least 50% of occasions completed
4	At least 75% of occasions completed
Decision 2: Inclusion of measurement occasions	
1	All (valid) occasions
2	Only occasions at which at least one negative emotion was experienced
Decision 3: ICC measurement unit	
1	Single measurements: ICC (3, 1)
2	Average of *k* measurements: ICC (3, *k*)
Decision 4: ICC agreement type	
1	Consistency
2	Absolute agreement
Decision 5: Treatment of negative ICCs	
1	Exclusion of negative ICCs
2	Setting negative ICCs to zero
3	Inclusion of negative ICCs as they are
Decision 6: Transformation of ICCs	
1	No transformation of ICCs
2	Fisher's *Z*-transformation of ICCs
Total number of combinations: 192	

Note. Decision 1 resulted in the following absolute values: 10, 21, 42, 63 occasions (Study 1); 3, 5, 10, 15 days (Study 2), and 14, 28, 56, 84 occasions (Study 3). The total number of combinations is the product of the number of options in the crossed decisions. The decision on the inclusion of measurement occasions was not preregistered, but resulted from the observation that negative emotions were experienced in only 22% of the measurement occasions in Study 1. ICC = intraclass correlation.

75% as a higher threshold due to generally high compliance rates in AA studies offering monetary incentives (Ottenstein & Werner, 2022).

Inclusion of Measurement Occasions

The ICC is typically calculated across all measurement occasions for each participant. However, when participants repeatedly rate the momentary intensity of discrete emotions such as anger or embarrassment—affective states which have an onset, a certain duration, and may be absent at times (Frijda, 1993; Lischetzke & Könen, 2022)—the question arises as how to handle occasions at which no negative emotions are experienced. In a study measuring NED through responses to standardized stimuli, Nook et al. (2018) excluded participants who did not report experiencing emotions in at least 50% of the trials. This exclusion ensured that the NED index was representative of instances where negative emotions required differentiation. This rationale also applies to the exclusion at the level of individual measurement occasions rather than entire participants. Thus, in our multiverse analysis, we explored the decision to calculate the ICC either across all measurement occasions, regardless of the presence of negative emotions, or only for occasions at which at least one negative emotion was experienced (*occasions with negative emotions*).[3]

ICC Measurement Unit

The calculation of the ICC involves partitioning the variance in emotion ratings into three components: variance due to differences between occasions, variance due to differences between emotions, and residual variance (Field, 2005; Schreuder et al., 2022). Different ICC specifications correspond to different formulae, which determine how these variance components are treated in the calculation of the ICC (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

The decision regarding the unit of measurement concerns whether the ICC (3) is computed based on single measurements (ICC [3, 1]) or on the average of *k* measurements (ICC [3, *k*]). The unit of measurement refers to the type of emotion ratings entered into the calculation (rather than to measurement occasions). Specifically, the emotion ratings either represent individual item scores or scores aggregated

3 This decision in the multiverse analysis was not preregistered. It was added to address the observation that only at 22% of the measurement occasions in Study 1 at least one negative emotion was reported.

across multiple items (e.g., emotion terms averaged within a broader emotion category). Conceptually, ICC (3, 1) is appropriate when using individual item scores, whereas ICC (3, *k*) is appropriate when using averaged scores across multiple items (Field, 2005; McGraw & Wong, 1996). Computationally, ICC (3, 1) includes residual variance in the denominator, while ICC (3, *k*) excludes this component, typically resulting in higher ICC estimates (Koo & Li, 2016). Consequently, this choice can influence NED scores. While some researchers used single measurements (e.g., Ottenstein & Lischetzke, 2020, Study 2), others have reported using the average of *k* measurements (e.g., Hoemann et al., 2023). However, this information is often not explicitly reported. We therefore include both approaches in our multiverse analysis.

ICC Agreement Type

Another decision researchers have to make is whether to compute ICCs as indices of absolute agreement or consistency (McGraw & Wong, 1996; Shrout & Fleiss, 1979). The absolute agreement ICC assesses whether emotion ratings are identical in their absolute value (i.e., equivalence). Computationally, the absolute agreement ICC incorporates variance attributable to systematic differences between emotions into the denominator. As a result, the absolute agreement ICC is affected by the absolute momentary emotion intensity, and considers emotions to be differentiated if their absolute values differ (Schreuder et al., 2022). In contrast, the consistency ICC excludes variance attributable to systematic differences between emotions from the denominator (Field, 2005). It therefore reflects the extent to which momentary emotion ratings covary over time (i.e., relative agreement). Under this specification, emotions are considered to be differentiated when they change independently over time (Schreuder et al., 2022).

Both types of ICCs are often highly correlated (Erbas et al., 2014; Lazarus & Fisher, 2021), but consistency ICCs tend to be larger than absolute agreement ICCs (Koo & Li, 2016), which may affect findings in NED research. Since only correlations between emotions are deemed relevant for NED (Feldman Barrett et al., 2001), consistency should be preferred (Erbas et al., 2014; Ottenstein & Lischetzke, 2020). However, absolute agreement has been applied in previous studies (e.g., Eckland et al., 2022; Mikhail et al., 2020; Nook et al., 2018; Willroth et al., 2020), so both options were included in our multiverse analysis.

Treatment of Negative ICCs

Although the ICC theoretically lies between 0 and 1 (Koo & Li, 2016), negative values can occur empirically (Giraudeau, 1996). Negative ICCs may result from random error or measurement issues, such as a lack of variability in emotion ratings (Erbas et al., 2019; Schreuder et al., 2022). However, they may also reflect a high level of emotion differentiation (Erbas et al., 2019; Thompson, Liu, et al., 2021). For instance, Thompson, Liu, et al. (2021) found that participants with negative ICC values exhibited similar response patterns as those with positive ICC values close to zero (i.e., high differentiation). Two common strategies have emerged in NED research to handle negative ICCs, each aligned with one of these interpretations (Thompson, Springstein, et al., 2021): (1) excluding them from analyses (e.g., Erbas et al., 2018; Kalokerinos et al., 2019; Lischetzke et al., 2021), as they are considered theoretically impossible (Giraudeau, 1996), or (2) setting them to zero (e.g., Hoemann et al., 2023; Liu et al., 2020), as they may indicate high differentiation (Thompson, Liu, et al., 2021). A third approach is to include negative ICCs, recognizing the potential validity of these values under certain conditions (e.g., Mikhail et al., 2020). Therefore, our multiverse analysis incorporated all three options for treating negative ICC values: exclusion, setting to zero, and inclusion as they are.

Transformation of ICCs

The final decision for NED indices is whether to transform the ICCs because they are usually not normally distributed (Erbas et al., 2019). Fisher's *Z*-transformation is often used to achieve normally distributed indices (e.g., Kalokerinos et al., 2019; Lischetzke et al., 2021; Liu et al., 2020), but some researchers either omit or do not report it. Our multiverse analysis included both Fisher's *Z*-transformation and no transformation.

The Present Research

The present research aimed to examine the robustness of empirical findings on the relationship between NED and adaptive outcomes (i.e., lower depression, reduced momentary stress reactivity) across typical decisions in data processing. Using a multiverse analysis approach, we investigated the robustness of results across 192 universes, which represent all possible combinations of six methodological decisions. In addition to assessing whether effects were robust across the multiverse, we examined whether particular decisions systematically influenced effect estimates. We focused on two analyses of interest: 1) the bivariate correlation between NED and depression, and 2) the buffering effect of NED on momentary stress reactivity, operationalized as a cross-level interaction in multilevel models. Data were drawn from three AA studies in which participants completed a baseline measure of depression and subsequently reported their emotional experiences, stress, and mood multiple times per day across 14 to 21 consecutive days. Separate multiverse analyses were conducted for each analysis of interest within each study. Study 1 was specifically designed to address the present research question (along with other unrelated research questions). Studies 2 and 3 were re-analyses of existing AA data sets (Lischetzke et al., 2021; Schmitt et al., 2024) and served to extend the investigation to samples with different design characteristics and higher proportions of measurement occasions involving negative emotions. Study 2 replicated the key variables of Study 1 in a daily diary design (morning and evening assessments). Study 3 further extended the buffering hypothesis by operationalizing stress reactivity via pleasant-unpleasant mood rather than calm-tense mood. Impairment in pleasant mood has likewise been

identified as an indicator of stress reactivity (Klaperski et al., 2013), suggesting that the theoretical rationale for a buffering effect of NED should generalize to this operationalization. Key methodological features of the three studies are summarized in Table S1 in the Online Supplement.

Method

The present research draws on data from three AA studies conducted in German samples. Below, we describe only those methodological aspects relevant to the present research. Study 1 is presented in greater detail, as it was specifically designed for the present investigation. A comprehensive codebook of all variables assessed in Study 1 is available on OSF (https://osf.io/jdesw/). Studies 2 and 3 are summarized more briefly (for a detailed description of Study 2, see Lischetzke et al., 2021; for Study 3, see Schmitt et al., 2024). Participants in these studies provided consent for the re-use of their data beyond the original research purposes.

Participants and Procedure

Study 1

Participants were recruited through advertisements (at the university, doctors' offices, and shops), social media, email distribution lists, and a press release. They had to be at least 18 years old and have a smartphone (Android or iOS). After registration, participants received detailed study and data protection information and provided informed consent before completing a 10-minute baseline questionnaire assessing depression.

The subsequent 21-day AA phase included four prompts per day. Participants chose between an early and a late schedule. In the early (late) schedule, prompts were sent at 9:00 a.m. (10:30 a.m.), 12:00 p.m. (2:00 p.m.), 4:30 p.m. (5:30 p.m.), and 9:00 p.m. (10:30 p.m.), expiring after 65 minutes. The prompts took about three minutes each, and assessed momentary mood, stress, and emotional experience. To address an unrelated research question, participants were randomly assigned to one of two experimental groups that differed in the response format (Likert scales vs. slider scales) for selected AA items. Of these items, only the calm-tense mood measure was relevant to the present research (i.e., the test of the buffering effect of NED on stress reactivity). Because it was unclear at the time of preregistration whether the two mood measures were psychometrically equivalent, we treated the two groups as separate samples. To remain consistent with this analytic strategy, we preregistered and retained this separation in the analyses examining the bivariate correlation between NED and depression.

After the AA phase, participants completed a 10-minute final online questionnaire. Data were collected using LimeSurvey (LimeSurvey GmbH, 2022) and SEMA³ (for the AA data; O'Brien et al., 2024) between April 2023 and February 2024. Compensation was based on AA compliance (additional to the online surveys): Participants earned €15 (€40) for at least 50% (80%) compliance. Psychology students could receive course credit instead. All participants were offered individualized feedback on their AA responses. The study was approved by the ethics committee of the Department of Psychology at the RPTU University Kaiserslautern-Landau (approval number LEK-439).

The final sample comprised 406 participants (318 female, 4 non-binary) with a mean age of 27.39 years (*SD* = 9.20, *Range* = 18-68), who provided 21,552 valid occasions. At 4,737 occasions (22%), participants reported experiencing at least one negative emotion, corresponding to an average of 11.67 such occasions per participant (*SD* = 12.27, *Range* = 0-76). Descriptive statistics for the two experimental groups are provided in Table S2 in the Online Supplement.

Study 2

Study 2 consisted of a baseline online survey assessing depression followed by a 21-day AA phase with two daily surveys: a morning survey assessing momentary mood and an evening survey assessing momentary mood, daily stress, and emotional experience. Data were collected in 2020. The final sample included 327 participants (243 women, mean age = 29.90 years, *SD* = 14.90, *Range* = 15-82) with a mean compliance of 18.58 days (*SD* = 3.20, *Range* = 1-21) out of 21 study days. In total, participants provided 11,876 valid measurement occasions, of which 5,920 were evening surveys. At least one negative emotion was reported in 5,755 evening surveys (97.2%). On average, participants reported at least one negative emotion on 17.6 days (*SD* = 3.55, *Range* = 1-21).

Study 3

Study 3 consisted of a baseline online questionnaire assessing depression followed by a 14-day AA phase with eight prompts per day (112 planned occasions) assessing mood, stress, and negative emotion intensity. Data were collected in 2021. The final sample included 163 participants (128 female, 3 non-binary; mean age = 30.87 years, *SD* = 9.21, *Range* = 19-64), who provided 12,400 valid measurement occasions. On average, participants completed 76.71 occasions (*SD* = 29.61, *Range* = 1-111). At least one negative emotion was reported at 12,122 occasions (97.8%), corresponding to an average of 74.37 such occasions per participant (*SD* = 29.08, *Range* = 1-111).

Measures

Within-Person (Momentary) Measures

Momentary Stress. In Study 1, momentary stress was assessed using a single item (similar to Erbas et al., 2018) with a unipolar 5-point Likert scale ranging from 1 (*not stressed at all*) to 5 (*very stressed)*. It read, "How stressed have you felt since you got up?" at the first prompt per day and "How stressed have you felt since you completed the last survey?" at subsequent prompts. In Study 2, daily stress was assessed retrospectively in the evening survey using a

single item on a slider scale (0 = *not at all* to 100 = *very much*). In Study 3, momentary stress over the past hour was assessed with a single item on a 101-point slider scale from 0 (*not at all*) to 100 (*very much*).

Momentary Mood. In all studies, momentary mood was assessed with an adapted short version of the Multidimensional Mood Questionnaire (Steyer et al., 1997). In Studies 1 and 2, calm-tense mood was measured using two bipolar items (*tense–relaxed, calm–restless* [reverse-poled]) with the word pairs as verbal anchors. In Study 1, response format depended on experimental group (5-point Likert vs. 101-point slider scale). In Study 2, both items were rated on a 101-point slider scale. In Study 3, pleasant-unpleasant mood was assessed using four bipolar items (*unwell-well, bad-good, unsatisfied-satisfied, unhappy-happy*) on a 101-point slider scale with verbal labels on each pole.

Mood items were averaged per occasion so that higher scores represented higher momentary calmness (Studies 1 and 2) or more pleasant mood (Study 3). Within-person reliability estimates (Lai, 2021) were α = .76 (Likert scales group) and α = .75 (slider scales group) in Study 1, α = .76 in Study 2, and ω = .92 in Study 3.

Momentary Negative Emotion Intensity. In Study 1, participants first indicated whether they had experienced any negative emotion since getting up (first prompt of the day) or since the last prompt. If so, they rated the intensity of 10 discrete negative emotions (fear, anger, frustration, sadness, embarrassment, envy, remorse, disgust, pity, contempt) using a 5-point Likert scale (0 = *not at all*, 1 = *very weak*, 2 = *rather weak*, 3 = *rather strong*, 4 = *very strong*). The items were derived from the German version of the Geneva Emotion Wheel (Sacharin et al., 2012).

In Study 2, participants rated the daily intensity of eight negative emotions (anger, boredom, disappointment, embarrassment/shame, fear, loneliness, regret, sadness) in the evening survey on a slider scale (0 = *not at all,* 100 = *very intense*).

In Study 3, negative emotion intensity during the past hour was assessed at each prompt using 15 individual emotion terms, each presented in a separate item and rated separately on a 101-point slider scale (0 = *not at all*, 100 = *very much*). The items reflected five broader emotion categories (anger: angry, irritated, annoyed; sadness: sad, downhearted, unhappy; fear: terrified, fearful, worried; shame: ashamed, humiliated, disgraced; guilt: guilty, repentant, I had a bad conscience)[4]. These categories were used for conceptual organization only and did not affect item presentation. The individual emotion intensity ratings were used to calculate the NED index (see Between-Person [Trait] Measures).

Between-Person (Trait) Measures

Negative Emotion Differentiation. In all studies, NED was operationalized as the ICC of momentary negative emotion intensity ratings across occasions within each participant. In Study 1, emotion intensity ratings were set to zero when participants denied experiencing any negative emotion at a given prompt (relevant only for universes including all occasions). The ICC specification (ICC [3, 1] vs. ICC [3, *k*]; consistency vs. absolute agreement), the treatment of negative ICCs, and transformation of ICCs varied across universes. ICC values were subtracted from 1 so that higher values reflected greater NED.

Depression. In Study 1, depression was measured using the 7-item depression subscale of the German short version of the Depression Anxiety Stress Scale (DASS-21; Nilges & Essau, 2021). Participants indicated the extent to which each statement (e.g., "I felt down-hearted and blue") applied to them in the past month using a 4-point Likert scale (0 = *did not apply to me at all*, 1 = *applied to me to some degree or some of the time*, 2 = *applied to me to a considerable degree or a good part of the time*, 3 = *applied to me very much or most of the time*). In Studies 2 and 3, depression was assessed in reference to the past two weeks using nine items of the depression scale of the Patient Health Questionnaire (Gräfe et al., 2004; Spitzer et al., 1999) on a 4-point Likert scale ranging from 0 (*not at all*) to 3 (*almost every day*) in Study 2, and from 1 (*not at all*) to 4 (*almost every day*) in Study 3. The depression items were aggregated to a mean score, such that higher scores indicated greater depression. McDonald's ω values were .88 (Study 1), .83 (Study 2), and .84 (Study 3).

Sample Size Considerations

Given the multilevel structure of AA data, sample size planning for Study 1 addressed both Level 1 (measurement occasions) and Level 2 (participants). Because Study 1 was part of a larger project, the planned number of measurement occasions was based on a separate research question involving multilevel models with fixed effects at Level 1. Anticipating 80% compliance (Ottenstein & Werner, 2022; Wrzus & Neubauer, 2023), we planned 84 occasions per participant. For the Level 2 sample size, we based our calculation on the expected cross-level interaction effect central to the present research, using effect size estimates from Lischetzke et al. (2021)[5]. An a priori power analysis using

4 This measurement structure is conceptually similar to that used in prior research examining between-category and within-category emotion differentiation (Erbas et al., 2019). It also aligns with several other studies that used multiple emotion terms per discrete emotion category and calculated a single overall NED index across all items (e.g., Emery et al., 2014; Starr et al., 2020; Willroth et al., 2020). Importantly, Erbas et al. (2019) found that an overall NED index (calculated across all emotion items) was more strongly associated with well-being outcomes than either within- or between-category indices. This finding supports the validity of overall NED scores derived from items sets that include multiple items per emotion category.

5 Initially, we had planned to use only the data from Study 1, which is why we conducted the power analysis with the data from Study 2. Studies 2 and 3 were added later to address limitations of Study 1.

Monte Carlo simulation (Bolger & Laurenceau, 2013) indicated that 165 participants with 65 measurement occasions each (reflecting ~80% compliance) would provide a power of .80. Because Study 1 included two experimental groups analyzed separately and we anticipated a dropout rate of approximately 10% prior to the AA phase (cf. Wrzus & Neubauer, 2023), we aimed to recruit at least 367 participants.

For Study 3, which operationalized stress reactivity via pleasant mood impairment, we conducted an additional Monte Carlo power analysis based on an effect size estimate derived from Lischetzke et al. (2021). Assuming 165 participants (consistent with Study 1) and an average of 76 measurement occasions per participant (the observed mean compliance in Study 3), the estimated power was 84%.

For the association between NED and depression, we expected a small to moderate correlation (r = -.20). A power analysis with G*Power (Version 3.1.9.7; Faul et al., 2007) indicated a power of 83% with 165 participants. Further details are provided in the preregistration (https://osf.io/jdesw/).

Data Preparation Strategy

Prior to implementing the multiverse specifications, we performed occasion-level data cleaning by excluding uncompleted measurement occasions and those flagged for careless responding. Specifically, we screened measurement occasions for careless responding by examining response consistency between semantic antonyms (Meade & Craig, 2012), following the procedure by Hasselhorn et al. (2022). Responses were classified as inconsistent when participants endorsed opposite extremes of the same bipolar mood dimension (e.g., feeling very tense and very relaxed at the same time) at a single occasion. This was operationalized as responses in the negative and positive extremes of the respective scale (Likert scales: -2 and +2; slider scales: ⩽ -30 and ⩾ +30 for scales ranging from -50 to +50, or ⩽ 20 and ⩾ 80 for scales ranging from 0 to 100, respectively). Occasions with inconsistent responses were excluded from the analyses. Information on participant enrollment and exclusions (for Study 1) and on the number of completed and excluded occasions (for all studies, e.g., due to careless responding) is reported in the Online Supplement (in the Additional Information on Data Preparation section).

Subsequent data preparation followed the multiverse specifications (e.g., compliance thresholds). In Study 1, all steps were conducted separately for the two experimental groups, resulting in four multiverse analyses (two analyses of interest per group). Participants were excluded if they (a) showed zero variance in emotion ratings, (b) had fewer than three occasions available for ICC calculation, (c) obtained an ICC of 1 (perfect agreement), or (d) obtained an ICC of ⩽ -1. The latter exclusion ensured comparability across universes with and without Fisher's Z-transformation, since transformed ICC values approach plus infinity (minus infinity) as when ICCs approach 1 (-1), causing estimation problems.

For analyses of mood change (see Data Analysis), we calculated lagged calmness and lagged pleasant mood variables reflecting the score at the previous measurement occasion. Lagged values were set to missing for the first occasion of each day (to model within-day change only) and when the preceding occasion had been classified as careless.

Finally, cases with missing data were excluded as required by the respective analyses (e.g., occasions with missing lagged mood scores for the buffering effect of NED).

Data Analysis

Analyses of Interest

All data preparation and analysis were performed in R (Version 4.5.0; R Core Team, 2025). For the first analysis of interest, we tested the correlation between NED and depression with a one-tailed test (α = .05) consistent with our directed hypothesis. Correlations were computed using the R package *psych* (Revelle, 2022).

For the second analysis of interest, we tested whether NED buffered the within-person association between stress and mood using a two-level model with random slopes. At Level 1, momentary stress and lagged mood predicted current mood; at Level 2, NED and the cross-level interaction between NED and stress were included. Including lagged mood allowed us to model the change in momentary mood, consistent with previous studies (Lischetzke et al., 2021; Starr et al., 2020). In Studies 1 and 2, calmness served as the outcome variable; in Study 3, pleasant mood was used instead. The two-level model predicting momentary mood for person i at occasion t was:

$$\begin{aligned} \text{Level 1: } \text{mood}_{ti} &= \pi_{0i} + \pi_{1i} * \text{stress}_{ti} \\ &\quad + \pi_{2i} * \text{mood}_{(t-1)i} + e_{ti} \\ \text{Level 2: } \pi_{0i} &= \beta_{00} + \beta_{01} * \text{NED}_{i} + r_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11} * \text{NED}_{i} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \end{aligned} \quad (1)$$

Level-1 predictors were person-mean centered, and NED was grand-mean centered (Wang & Maxwell, 2015). To facilitate model convergence, variables assessed on slider scales were rescaled to a range of -0.50 to +0.50 (Study 1) or 0 to 1 (Studies 2 and 3). Models were estimated with restricted maximum likelihood using *lme4* (Bates et al., 2015). We used the default optimizer and switched to *bobyqa* when convergence issues occurred. The cross-level interaction was tested with *lmerTest* (Kuznetsova et al., 2017), using a two-tailed test. We calculated 95% confidence intervals using the Satterthwaite-approximated degrees of freedom. As an effect size, we calculated the proportion of within-person outcome variance explained by the cross-level interaction (Rights & Sterba, 2020) using *r2mlm* (Shaw et al., 2023). To this end, we determined the difference in the proportion of within-person variance explained by the fixed Level-1 effects ($\Delta R_w^{2(f1)}$) between a model including and a model excluding the cross-level interaction.

Multiverse Analyses

After running the analyses of interest across all universes, we explored the impact of data preparation decisions on the effect estimates in three ways:

1. *Robustness*: We checked if the effects were statistically significant in the expected direction and relatively stable in magnitude across universes.
2. *Patterns*: We examined how the methodological decisions influenced the estimates by assessing whether specific choices tended to weaken or strengthen the observed effects. For each decision, we compared the distribution of effect estimates associated with each option to the overall median effect. We evaluated both the direction of the shift (i.e., whether the distribution was skewed toward stronger or weaker effects relative to the median) and the magnitude of this trend. Based on our hypotheses, we expected a negative correlation between NED and depression, and a positive cross-level interaction between NED and stress in predicting mood change. Accordingly, effects were considered strengthened when correlation estimates (NED-depression) were more strongly negative or when cross-level interaction estimates (NED x stress) were more strongly positive, and weakened when estimates shifted in the opposite direction. Importantly, we interpreted trends at the distributional level, not based on individual effect estimates. That is, although some individual universes for a given decision option may produce smaller or even opposite effects, the same option could still be associated with a general tendency toward stronger effects when considering the overall distribution.
   To quantify the strength of a trend, we classified it as weak, medium, or strong when the proportion of estimates on one side of the overall median was at least 1.2, 2, or 3 times greater than the proportion on the other side. For example, if 75% of the estimates associated with a given option fell below the overall median and 25% above, we classified this as a strong trend towards more negative (or less positive) effects.
3. *Influence of decisions*: To quantify the unique impact of each decision on the effect estimates, we applied a leave-one-out regression approach. The dummy-coded decisions predicted effect estimates (separately for each multiverse analysis) in a series of linear regressions: We first estimated a full model including all decisions and then iteratively left out one decision at a time. The resulting change in explained variance (ΔR^2) indicated the variance attributable to each decision. As this analysis was exploratory, we did not correct for multiple testing and refrain from reporting *p*-values.

Sensitivity Analyses: Controlling for Mean Negative Emotion Intensity

Given prior meta-analytic work emphasizing the importance of examining whether emotion dynamics constructs such as NED explain variance in psychological outcomes beyond mean emotion intensity (Dejonckheere et al., 2019), we conducted sensitivity analyses in which we re-ran the multiverse analyses while controlling for person-level mean negative emotion intensity.

Exploratory Analyses: Addressing Power Issues

In some universes, participant exclusions (due to a negative ICC or low compliance) reduced the Level-2 sample size and, consequently, the statistical power of the significance tests. Although such reductions reflect common research practice, differences in power across universes may confound the effects of data preparation decisions. To isolate the impact of these decisions from variation in power, we conducted exploratory resampling analyses: For universes with reduced Level-2 sample sizes, we resampled 165 participants with replacement (retaining their valid measurement occasions after data preparation) and re-ran the analysis. We used 165 participants for both analyses of interest, as the power analyses indicated sufficient power with this sample size. For Study 2, no resampling analyses were conducted, as the Level-2 sample size was sufficient in all universes.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Study 1's design, hypotheses and data analysis plan were preregistered before data collection started. Studies 2 and 3 were re-analyses of existing data. The preregistration of Study 1, study materials, data, data analysis code (including R packages and versions) and data sets with results of the multiverse analyses for all studies can be accessed via OSF (https://osf.io/jdesw/).

Results

We summarize the multiverse results across the three studies, with detailed information provided in Tables 2 and 3. For each multiverse analysis, these tables report the number of significant effects, the range of effect sizes, the sample sizes per universe, the number of negative ICCs, and the most influential decision. Descriptive statistics and correlations for all within- and between-person measures are presented in Tables S3 and S4 in the Online Supplement. Distributions of NED across the universes are depicted in Figures S1 to S4 (Online Supplement).

Multiverse Analyses on the Bivariate Correlation Between NED and Depression

Figures 1 to 4 display the multiverse analysis results for the correlation between NED and depression across the three studies. In each figure, the upper panel shows the

correlation estimates with 95% confidence intervals across all 192 universes (sorted in ascending order), with significant effects highlighted in cyan. The lower panel indicates the decision options selected in each universe; for example, a dot at the 75% compliance threshold indicates that this threshold was chosen.

Correlation estimates varied in both magnitude and statistical significance across data preparation decisions. The proportion of correlations that were significantly negative (as theoretically expected) ranged from 8% in Study 2 to 95% in Study 1 (Likert scales group). Accordingly, the range of the correlation estimates varied across the three studies, with the smallest effect sizes observed in Study 2 (ranging from -.13 to .04) and the largest effect sizes observed in the Likert scales group in Study 1 (ranging from -.53 to -.08).

Notably, more than twice as many universes yielded significant correlations between NED and depression in the Likert scales group compared to the slider scales group in Study 1[6]. Overall, the association between NED and depression was not robust in any study, as neither statistical significance nor effect size was stable across universes.

The prevalence of negative ICCs[7] (prior to their treatment) also differed across studies. The Likert scales group in Study 1 showed the highest proportion (range from 0% to 18%), whereas Study 2 exhibited negative ICCs in all universes (proportions ranging from 3% to 6%), and Study 3 showed none (see Table 2).

Overall, there was no consistent pattern in how the decisions affected the distribution of effect estimates (for detailed results, see Table 4)[8]. The only decision with a consistent directional impact across studies was the choice between consistency and absolute agreement, with consistency generally yielding larger effect sizes. The inclusion of all occasions versus only occasions with negative emotions influenced the correlation estimates only in Study 1, where including all occasions produced larger effects. In Studies 2 and 3, where negative emotions were reported on nearly all occasions (97% and 98%, respectively), this decision had no discernable impact.

The decision exerting the strongest impact on the correlation estimates varied by study (for details, see Table S6 in the Online Supplement). In Study 1, the compliance threshold was most influential in the Likert scales group, while the inclusion of occasions had the greatest impact in the slider scales group; in Studies 2 and 3, the consistency versus absolute agreement specification showed the strongest influence.

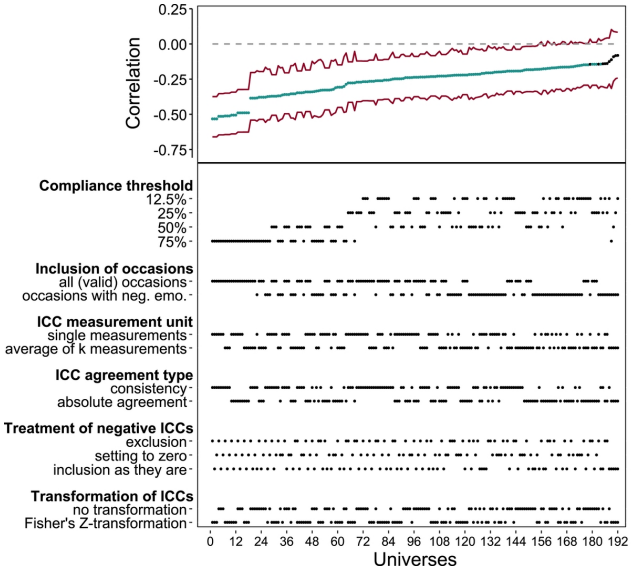

Figure 1. Results from the Multiverse Analysis on the Bivariate Correlation Between Negative Emotion Differentiation and Depression in the Likert Scales Group in Study 1

Note. The upper panel of the figure depicts the correlations and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant correlations are indicated in cyan, non-significant correlations are black. The dots aligned vertically below the correlations in the lower panel of the figure indicate which options of the decisions corresponded to the correlation in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

Multiverse Analyses on the Buffering Effect of NED on Within-Person Stress Reactivity

Figures 5 to 8 show the multiverse analysis results for the buffering effect of NED on within-person stress reactivity. Across studies, the proportion of significantly positive

6 Importantly, although the response format differed between groups for some items, the emotion items used to compute the NED indices were assessed using the same Likert response format in both groups. Of the items that differed in response format, only calm-tense mood was relevant to the analysis of the buffering effect of NED on stress reactivity, whereas none were relevant to the analysis of the correlation between NED and depression. A randomization check revealed no significant differences between the two groups in demographic characteristics, depression levels, mean emotion levels, NED levels, compliance rates, or the number of occasions with reported negative emotions. Subsequent analyses of the psychometric properties of the Likert and slider scales in this dataset suggest that the two formats do not differ meaningfully (Vollbracht et al., 2026). Therefore, there is no clear empirical basis for the observed difference in the distribution of correlation estimates between groups. This pattern highlights the potential sensitivity of multiverse results to seemingly peripheral design features (such as response formats for other parts of an AA survey). This underscores the importance of carefully considering and transparently reporting assessment methods in AA studies, and it highlights the need for future research to systematically investigate how such methodological variations influence NED estimation and related constructs.

7 To explore the relationship between the average number of measurement occasions per participant and the number of negative ICCs, we calculated their correlation across the universes. The results of these analyses for the multiverse analyses in all studies are summarized in Table S5 in the Online Supplement.

8 We examined these patterns using effect estimate histograms. The histograms for both analyses of interest are displayed in Figures S5 to S12 in the Online Supplement.

Table 2. Overview of the Multiverse Analysis Results on the Correlation between Negative Emotion Differentiation and Depression

Results	Study 1		Study 2	Study 3
	Likert Scales Group	Slider Scales Group		
Negative correlations ($N_{significant}$)	192 (182)	168 (68)	134 (15)	162 (72)
Positive correlations ($N_{significant}$)	0 (0)	24 (0)	58 (0)	30 (0)
Range of correlations	-.53 – -.08	-.30 – .08	-.13 – .04	-.24 – .09
Range of $N_{Level\ 2}$	75 – 168	87 – 176	274 – 325	86 – 154
Range of $N_{negative\ ICCs}$	0 – 27	0 – 17	10 – 17	0
Most influential decision	compliance (ΔR^2 = .61)	inclusion of occasions (ΔR^2 = .59)	ICC agreement type (ΔR^2 = .74)	ICC agreement type (ΔR^2 = .67)

Note. A negative correlation between negative emotion differentiation and depression was expected. In total, there were 192 universes in each multiverse analysis. The most influential decision was the decision that explained most variance in the effect estimates (ΔR^2). $N_{significant}$ = number of significant results. $N_{Level\ 2}$ = number of participants included in a given universe. $N_{negative\ ICCs}$ = number of negative ICCs in a given universe. ICC = intraclass correlation.

Table 3. Overview of the Multiverse Analysis Results on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity

Results	Study 1		Study 2	Study 3
	Likert Scales Group	Slider Scales Group		
Positive CLI estimates ($N_{significant}$)	95 (2)	32 (0)	192 (117)	192 (192)
Negative CLI estimates ($N_{significant}$)	97 (0)	160 (0)	0 (0)	0 (0)
Range of CLI estimates	-0.20 – 0.24	-0.05 – 0.02	0.03 – 0.30	0.16 – 0.80
Range of variance explained by CLI	-0.1% – 0.3%	-0.04% – 0.3%	0% – 0.3%	0.6% – 2%
Range of $N_{Level\ 2}$	75 – 168	87 – 176	274 – 325	86 – 154
Range of $N_{Level\ 1}$	3,600 – 6,261	4,185 – 6,731	5,060 – 5,731	6,650 – 9,012
Range of $N_{negative\ ICCs}$	0 – 27	0 – 17	10 – 17	0
Most influential decision	inclusion of occasions (ΔR^2 = .71)	ICC measurement unit (ΔR^2 = .19)	ICC measurement unit (ΔR^2 = .88)	transformation of ICCs (ΔR^2 = .47)

Note. A positive cross-level interaction (CLI) was expected. In total, there were 192 universes in the multiverse analysis. The most influential decision was the decision that explained most variance in the effect estimates (ΔR^2). $N_{significant}$ = number of significant results. $N_{Level\ 2}$ = number of participants included in a given universe. $N_{Level\ 1}$ = number of measurement occasions included in a given universe. $N_{negative\ ICCs}$ = number of negative ICCs in a given universe. ICC = intraclass correlation.

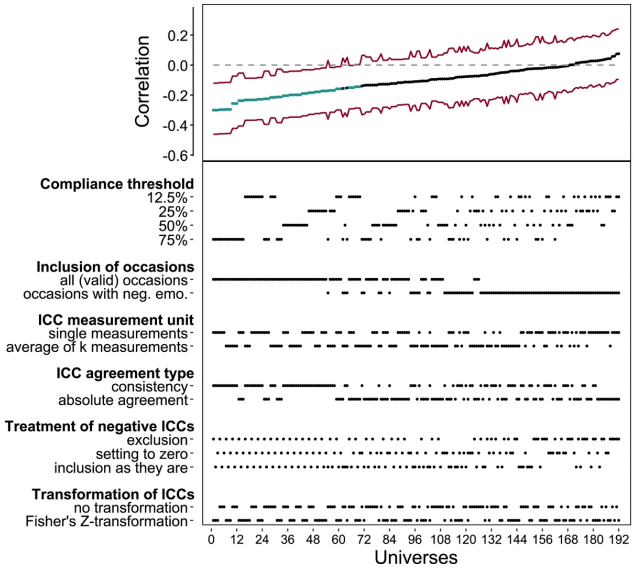

Figure 2. Results from the Multiverse Analysis on the Bivariate Correlation Between Negative Emotion Differentiation and Depression in the Slider Scales Group in Study 1

Note. The upper panel of the figure depicts the correlations and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant correlations are indicated in cyan, non-significant correlations are black. The dots aligned vertically below the correlations in the lower panel of the figure indicate which options of the decisions corresponded to the correlation in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

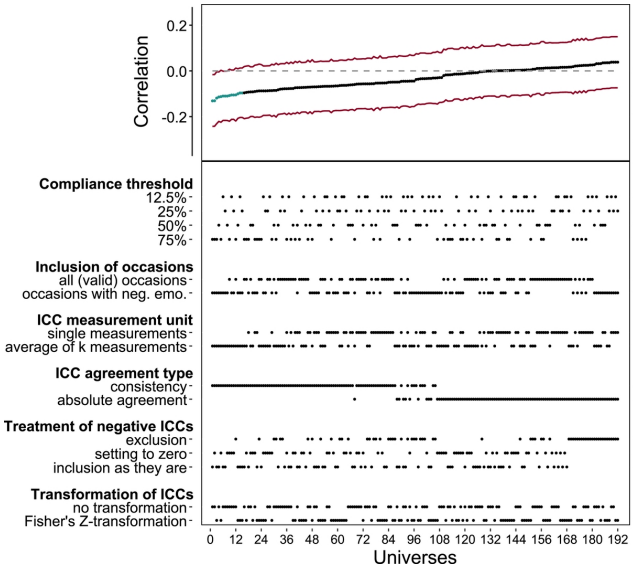

Figure 3. Results from the Multiverse Analysis on the Bivariate Correlation Between Negative Emotion Differentiation and Depression in Study 2

Note. The upper panel of the figure depicts the correlations and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant correlations are indicated in cyan, non-significant correlations are black. The dots aligned vertically below the correlations in the lower panel of the figure indicate which options of the decisions corresponded to the correlation in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

cross-level interaction estimates (as theoretically expected) ranged from 0% (slider scales group in Study 1) to 100% (Study 3). Accordingly, both the magnitude of the cross-level interaction estimates and the proportion of within-person variance explained varied across studies.

The largest positive effects were observed in Study 3, where cross-level interaction estimates ranged from 0.16 to 0.80 and explained between 0.6% and 2% of within-person variance. In contrast, Study 1 yielded mostly non-significant negative cross-level interaction estimates. Thus, a robust buffering effect of NED on within-person stress reactivity was observed only in Study 3, in which stress reactivity was operationalized as pleasant mood impairment rather than reduced calmness.

The highest prevalence of negative ICCs (prior to treatment) was observed in the Likert scales group of Study 1 (proportion ranging from 0% to 18%), whereas Study 2 showed negative ICCs in all universes (proportion ranging from 3% to 6%), and Study 3 showed none (see Table 3).

Overall, there was no consistent pattern in how the decisions affected the distribution of the cross-level interaction estimates (see Table 5).

The only consistent finding across the three studies was that the treatment of negative ICCs did not have a systematic impact on the cross-level interaction estimates. The decision concerning the inclusion of occasions affected the estimates only in Study 1. Including all occasions in the ICC calculation weakened the estimated buffering effect, whereas calculating ICCs using only occasions with negative emotions yielded stronger interaction effect estimates. This decision did not affect the results in Studies 2 and 3.

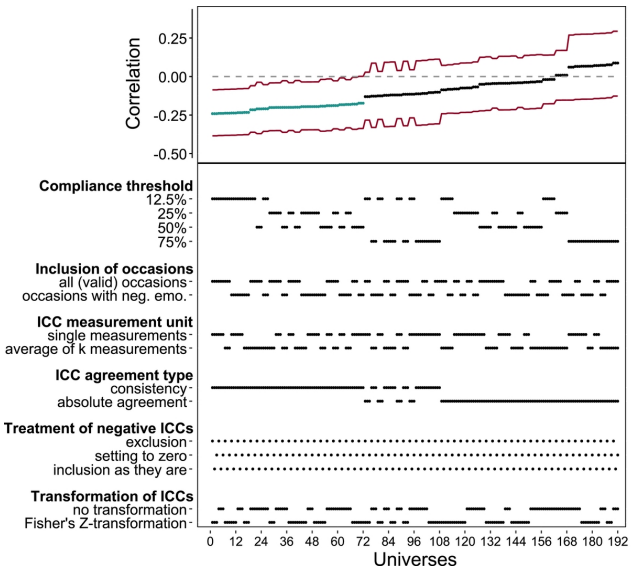

Figure 4. Results from the Multiverse Analysis on the Bivariate Correlation Between Negative Emotion Differentiation and Depression in Study 3

Note. The upper panel of the figure depicts the correlations and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant correlations are indicated in cyan, non-significant correlations are black. The dots aligned vertically below the correlations in the lower panel of the figure indicate which options of the decisions corresponded to the correlation in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

Table 4. Effects of the Data Preparation Decisions on the Correlation Between Negative Emotion Differentiation and Depression

Decision	Option	Study 1		Study 2	Study 3
		Likert Scales Group	Slider Scales Group		
Compliance threshold	12.5%	↓↓↓	=	=	↑↑↑
	25%	↓↓	↓	=	=
	50%	=	=	=	=
	75%	↑↑↑	↑	↑	↓↓↓
Inclusion of occasions	All (valid) occasions	↑↑	↑↑↑	=	=
	Occasions with negative emotions	↓↓	↓↓↓	=	=
ICC measurement unit	Single measurements	↑	=	↓	=
	Average of *k* measurements	↓	=	↑	=
ICC agreement type	Consistency	↑	↑	↑↑↑	↑↑↑
	Absolute agreement	↓	↓	↓↓↓	↓↓↓
Treatment of negative ICCs	Exclusion	=	↓	↓	=
	Setting to zero	=	=	=	=
	Inclusion as they are	=	↑	↑	=
Transformation of ICCs	No transformation	↓	=	=	=
	Fisher's *Z*-transformation	↑	=	=	=

Note. ↑ (↓) indicates that the correlations tend to be more negative (less negative). The number of arrows indicates the strength of the trend (1 = weak, 2 = medium, 3 = strong). = indicates that approximately equal proportions of the effect estimates are on either side of the median, that is, there is no clear trend. Occasions with negative emotions = measurement occasions with experience of negative emotion. ICC = intraclass correlation.

The most influential decisions differed by study. Occasion inclusion exerted the strongest influence in Study 1 (Likert scales group; ΔR^2 = .71), the ICC measurement unit was most influential in Study 1 (slider scales group; ΔR^2 = .19) and in Study 2 (ΔR^2 = .88), and the transformation of ICCs had the largest impact in Study 3 (ΔR^2 = .47). The influence of the remaining decisions is reported in Table S7 in the Online Supplement[9].

Sensitivity Analyses: Controlling for Mean Negative Emotion Intensity

As sensitivity analyses, we repeated the multiverse analyses for both analyses of interest while controlling for mean negative emotion intensity. Detailed results are provided in the Online Supplement (Tables S10 to S17, and Figures S13 to S28). For the correlation between NED and depression, the effect sizes decreased when controlling for mean negative emotion intensity, resulting in fewer or no significant universes in the expected direction. In Study 2, all correlations between NED and depression were unexpectedly positive, with 40% being significant. An exception emerged in the slider scales group in Study 1, where a larger proportion of universes yielded significant effects (47%) compared to the analyses without controlling for mean negative emotion intensity (35%), and the divergence in the number of significant correlations between experimental groups was reduced. Furthermore, the influence of specific methodological decisions changed. In Study 1, the decision effects disappeared in most cases, and in Studies 2 and 3, new effects emerged (see Table S11 in the Online Supplement). Notably, in Study 2, restricting the ICC calculation to occasions with negative emotion led to more positive correlations between NED and depression, whereas including all occasions led to less positive correlations (though none were negative, as expected). Moreover, the impact of the consistency versus absolute agreement specification on the correlations between NED and depres-

9 To examine potential interactions between decisions, we conducted additional regression analyses, each including a single two-way interaction term between decisions. For each interaction, we calculated the proportion of variance explained (ΔR^2) by comparing the interaction model to a model that included only the main effects. Overall, the additional variance explained in the effect estimates was close to zero. There were, however, two exceptions concerning the buffering effect of NED on stress reactivity (i.e., the cross-level interaction estimates). In Study 1, the interaction between the compliance threshold and the inclusion of occasions explained an additional 26% of variance in the slider scales group. In Study 3, the interaction between the ICC measurement unit and the ICC transformation explained an additional 24% of variance. Full results are reported in the Online Supplement (Tables S8 and S9).

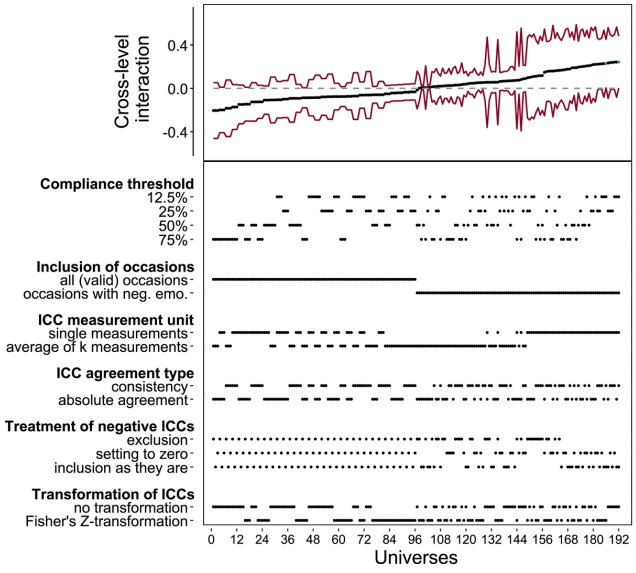

Figure 5. Results from the Multiverse Analysis on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity in the Likert Scales Group in Study 1

Note. The upper panel of the figure depicts the cross-level interaction (CLI) estimates and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant CLI estimates are indicated in cyan, non-significant CLI estimates are black. The dots aligned vertically below the CLI estimates in the lower panel of the figure indicate which options of the decisions corresponded to the CLI estimate in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

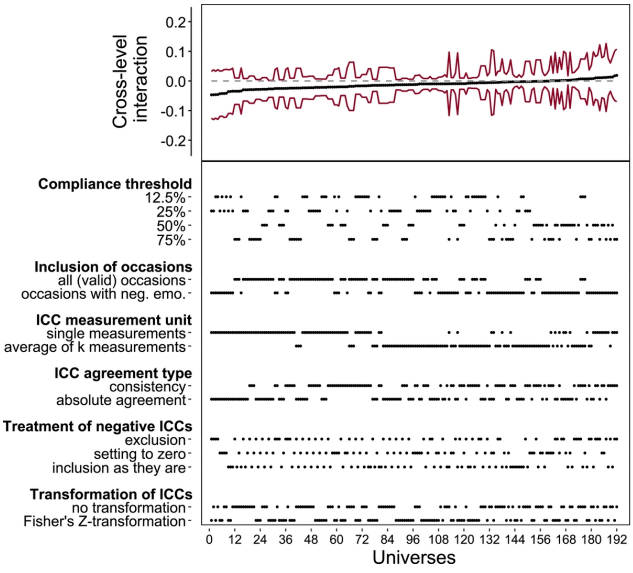

Figure 6. Results from the Multiverse Analysis on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity in the Slider Scales Group in Study 1

Note. The upper panel of the figure depicts the cross-level interaction (CLI) estimates and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant CLI estimates are indicated in cyan, non-significant CLI estimates are black. The dots aligned vertically below the CLI estimates in the lower panel of the figure indicate which options of the decisions corresponded to the CLI estimate in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

sion was weakened (Studies 2 and 3) or eliminated (Study 1). This pattern is in line with the fact that the consistency versus absolute agreement distinction is directly related to how variance attributable to mean-level differences between emotions is treated in the ICC computation.

In contrast, the results for the buffering effect of NED on stress reactivity were largely identical when controlling for mean emotion intensity. Overall, the conclusions regarding the robustness of the observed effects across methodological decisions remained unchanged: none of the effects were statistically significant in the expected direction across all universes in any data set, with the exception of the buffering effect in Study 3.

Exploratory Analyses: Addressing Power Issues

To account for reduced statistical power in universes with smaller Level-2 sample sizes, we conducted exploratory resampling analyses with 165 participants in Studies 1 and 3. Detailed results are reported in the Online Supplement (Tables S18 to S25, and Figures S29 to S40). Overall, the resampling analyses yielded patterns comparable to those observed in the original multiverse results for both analyses of interest. Although the proportion of significant estimates changed in most cases, the general conclusions remained unchanged. Some decision-related patterns became more or less pronounced, resulting in certain decisions emerging as newly influential or, in a few cases, no longer showing systematic effects (see Tables S19 and

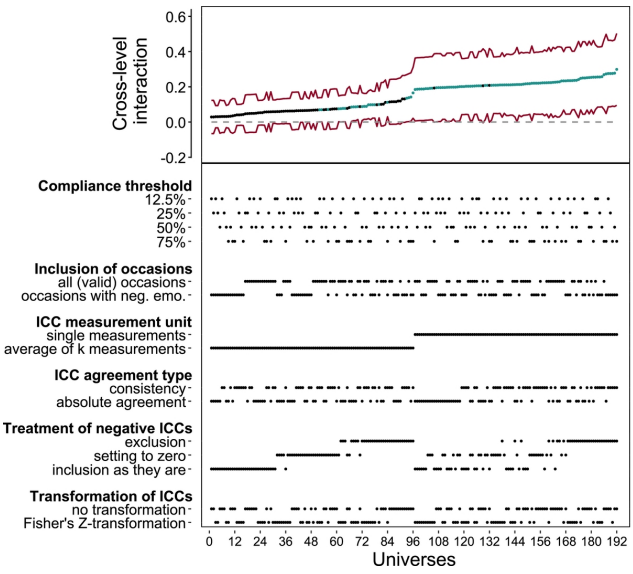

Figure 7. Results from the Multiverse Analysis on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity in Study 2

Note. The upper panel of the figure depicts the cross-level interaction (CLI) estimates and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant CLI estimates are indicated in cyan, non-significant CLI estimates are black. The dots aligned vertically below the CLI estimates in the lower panel of the figure indicate which options of the decisions corresponded to the CLI estimate in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

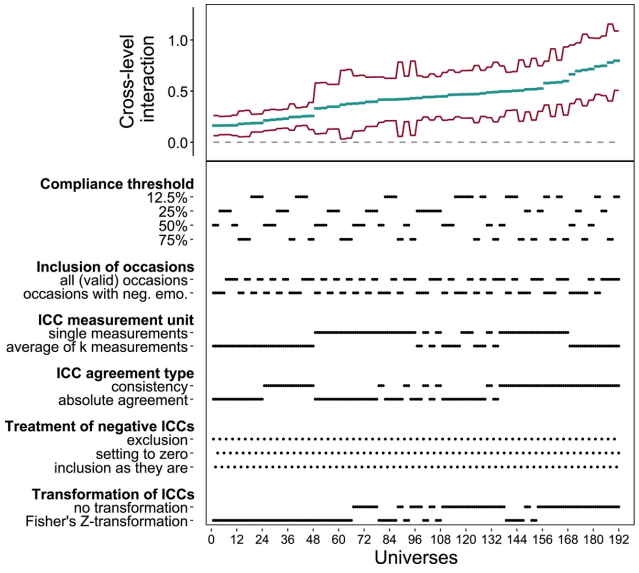

Figure 8. Results from the Multiverse Analysis on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity in Study 3

Note. The upper panel of the figure depicts the cross-level interaction (CLI) estimates and their 95% confidence intervals (dark red lines) in ascending order. The universes are plotted along the x-axis. Significant CLI estimates are indicated in cyan, non-significant CLI estimates are black. The dots aligned vertically below the CLI estimates in the lower panel of the figure indicate which options of the decisions corresponded to the CLI estimate in each universe. The dashed line marks zero. Occasions with neg. emo. = only measurement occasions with report of negative emotions included. ICC = intraclass correlation.

S23 in the Online Supplement). The identity of the most influential decisions remained largely stable across analyses. The only notable exception concerned the buffering effect of NED on stress reactivity in the slider scales group of Study 1, where the compliance threshold emerged as the most influential decision in the resampled analyses (see Tables S20 and S24 in the Online Supplement).

Discussion

The present research investigated whether and how methodological decisions affect empirical findings on the relationship between NED and adaptive outcomes. Using multiverse analyses across three AA studies, we evaluated the robustness of two analyses of interest: (a) the between-person association between NED and depression and (b) the buffering effect of NED on within-person stress reactivity. In addition, we conducted sensitivity analyses controlling for mean negative emotion intensity and, where necessary, resampling procedures to address potential power differences across universes.

Overall, neither analysis of interest was robust across methodological decisions. The only exception was the buffering effect of NED on stress reactivity in Study 3, in which stress reactivity was operationalized as pleasant mood impairment rather than reduced calmness. This finding is consistent with prior work (Klaperski et al., 2013) suggesting that pleasant mood impairment may capture stress reactivity more sensitively than calm mood.

Across studies and analyses, no single methodological decision consistently determined the pattern of results. Instead, the impact of specific decisions varied depending upon both the study and the analysis of interest. For example, the decision regarding occasion inclusion showed opposite effects across the two analyses of interest in Study 1: restricting the ICC calculation to occasions with negative emotions weakened the correlation between NED and depression but strengthened the buffering effect on stress reactivity. In Studies 2 and 3—where negative emotions were reported on nearly all occasions—this decision had no meaningful impact, except on the association between NED and depression when mean negative emotion intensity was controlled for in Study 2. Similarly, decisions concerning compliance thresholds, ICC measurement unit, and ICC transformation did not show uniform patterns across multiverse analyses.

Where systematic trends emerged, consistency ICCs generally yielded larger effect sizes and absolute agreement ICCs yielded smaller effect sizes. In contrast, the treatment of negative ICCs had little to no impact on the results. Across analyses, each decision—except for the treatment of negative ICCs—was identified as the most influential at least once, as reflected in the explanatory contribution to variation in effect estimates.

Notably, controlling for mean negative emotion intensity did not alter the results regarding the buffering effect of NED on within-person stress reactivity. This pattern is consistent with prior research. Of five AA studies examining the buffering effect (Lischetzke et al., 2021; Nook et al., 2021; Starr et al., 2017, Study 1 and Study 2; Starr et al., 2020), four reported a significant buffering effect that remained significant when controlling for mean negative emotion intensity (Lischetzke et al., 2021; Nook et al., 2021; Starr et al., 2017, Study 2; Starr et al., 2020). In contrast, controlling for mean negative emotion intensity attenuated effect sizes for the association between NED and depression in the present research and reduced the proportion of significant effects in most cases. This pattern aligns with previous findings. For example, a meta-analysis revealed that NED did not significantly predict depression when controlling for mean emotion intensity and explained little incremental variance beyond mean emotion intensity (Dejonckheere et al., 2019). Given this attenuation, the relationship between NED and depression appears particularly sensitive to data preparation decisions. Taken together, these findings suggest that NED may provide limited incremental predictive value for depression beyond mean negative emotion intensity, whereas it may play a more distinct role in buffering within-person stress reactivity.

Learnings from the Multiverse Analyses

The multiverse analyses reveal that, overall, empirical findings on NED's adaptive value are not robust across data preparation decisions, leading to variation in the effect estimates and their significance. As a result, substantive conclusions about NED may differ and sometimes even contradict each other. For example, NED could be associated with

Table 5. Effects of the Data Preparation Decisions on the Buffering Effect of Negative Emotion Differentiation on Within-Person Stress Reactivity

Decision	Option	Study 1		Study 2	Study 3
		Likert Scales Group	Slider Scales Group		
Compliance threshold	12.5%	=	=	=	↑
	25%	=	↓	=	=
	50%	=	↑	=	↓
	75%	=	=	=	=
Inclusion of occasions	All (valid) occasions	↓↓↓	↓↓↓	=	=
	Occasions with negative emotions	↑↑↑	↑↑↑	=	=
ICC measurement unit	Single measurements	=	↓↓↓	↑↑↑	=
	Average of *k* measurements	=	↑↑↑	↓↓↓	=
ICC agreement type	Consistency	=	↑	=	↑↑
	Absolute agreement	=	↓	=	↓↓
Treatment of negative ICCs	Exclusion	=	=	=	=
	Setting to zero	=	=	=	=
	Inclusion as they are	=	=	=	=
Transformation of ICCs	No transformation	=	=	=	↑↑↑
	Fisher's *Z*-transformation	=	=	=	↓↓↓

Note. ↑ (↓) indicates that the cross-level interaction estimates tend to be more positive (less positive). The number of arrows indicates the strength of the trend (1 = weak, 2 = medium, 3 = strong). = indicates that approximately equal proportions of the effect estimates are on either side of the median, that is, there is no clear trend. Occasions with negative emotions = measurement occasions with experience of negative emotion. ICC = intraclass correlation.

higher or lower levels of depression (e.g., in Study 2, when [not] controlling for mean emotion intensity). The variability associated with data preparation decisions could explain inconclusive empirical findings, such as opposite directions for the relationship between NED and depression in different samples (Eckland et al., 2022) or significant versus nonsignificant results at different time points (Liu et al., 2020) or with different scales (Lazarus & Fisher, 2021). Different results may arise from the same data preparation methods applied to different samples, time points or measures, as there appears to be no systematic pattern in how the decisions affect empirical results. The most influential decisions identified in the present research suggest that the variation is not only due to reduced power (e.g., from compliance decisions or negative ICC treatment), but stems also from how ICCs are calculated.

In general, the impact of the methodological decisions on the results of both analyses of interest increased when more participants, measurement occasions, or indices were affected. For instance, the treatment of negative ICCs was more relevant in Study 2 (in which each universe yielded at least 10 negative ICCs) than in Study 1 (where the proportion of universes without negative ICCs was higher). The higher proportion of universes without negative ICCs in Study 1 likely diluted the impact of the treatment of negative ICCs, as it would naturally have no effect in the absence of negative ICCs (as in Study 3). Similarly, the compliance thresholds had little to no impact in Study 2, where compliance was high, on average, and the extreme options led to a maximum difference of 34 participants (10% of the sample) between the universes with the most conservative (75%) and the most lenient (12.5%) option. As a result, the similar data sets of the universes probably produced similar effect estimates. Contrary, the impact of the compliance thresholds was stronger in Studies 1 and 3, where the difference pertained to a larger proportion of the sample. The same applies to the decision on which occasions to include for the ICC calculation, which only mattered in Study 1 where participants reported experiencing negative emotions at only a small number of occasions (22%). In contrast, the proportion of such occasions was very high in Studies 2 and 3. Thus, the universes that included all occasions and those that used only the occasions with negative emotions were almost identical and differed by only few occasions, such that the decision did not affect the results. An exception was found in Study 2, when we controlled for mean negative emotion intensity in the association between NED and depression. Here, the decision affected the estimates, which were all unexpectedly positive.

Finally, two aspects of interpreting multiverse analyses should be emphasized: (1) The results should be viewed in terms of *possibility*, not *probability*. A probabilistic interpretation assumes that all universes are equally likely to be correct, meaning that more frequent results are considered more likely. However, this is inaccurate because the universes are not randomly selected from all possible reasonable decisions, and the analyses are not statistically independent from each other (Hall et al., 2022; Simonsohn et al., 2020). In contrast, a possibilistic interpretation means that each result shows what is a possible result based on reasonable choices, without implying how likely it is (Hall et al., 2022). Therefore, we did not aggregate the results into a median effect and test its significance, as suggested by Simonsohn et al. (2020), as this could misrepresent the true variety of outcomes.

(2) We cannot provide specific recommendations for data preparation decisions based on our multiverse analyses. The purpose of such analyses is to assess the robustness of empirical findings, not to determine which result is true or which option should be chosen (Harder, 2020). However, general recommendations for dealing with researcher degrees of freedom in NED research are discussed in the Conclusion and Recommendations section.

Limitations

The conclusions of our multiverse analysis are subject to several limitations. Although a multiverse analysis systematically examines the robustness of empirical findings against somewhat arbitrary data preparation decisions, it is still subject to researcher judgment. Researchers may disagree about the most reasonable options (Simonsohn et al., 2020). For example, excluding participants due to low compliance is common in AA studies (e.g., Weermeijer et al., 2022), but the chosen thresholds can be arbitrary. In our case, a strict compliance threshold of 75% or limiting the analysis to four specific thresholds might be questioned. However, our aim was not to identify the most reasonable threshold, but to examine whether and how such decisions affect empirical findings. Therefore, we focused on a wide range of compliance with a small number of thresholds to ensure interpretability. For a more comprehensive multiverse analysis on compliance thresholds in AA studies, see Weermeijer et al. (2022).

Furthermore, the decisions examined are not exhaustive, as there are further decisions pertaining to study design and data analysis. For example, we did not examine factors relating to data analysis like centering of the predictors (Hamaker & Muthén, 2020) or accounting for careless responding. Instead, we limited the scope to the most relevant data preparation decisions for NED to avoid an even more complex analysis.

Relatedly, an important aspect that may have contributed to the inconsistent results is that the three studies differed in key methodological features, such as the emotion item set, sampling frequency, response format, and the reference timeframes for assessing state variables. For instance, Thompson, Springstein, and Boden (2021) argued that the number and type of emotion items—both of which varied across the present studies—is "the most critical issue in designing studies on differentiation" (p. 4). A recent NED study examined the role of the emotion terms by employing item sets that differed by one item, and reported mixed results on the robustness of NED's relationships with emotion regulation variables (Kalokerinos et al., 2019). This suggests that differences in the emotion terms used may partly explain the inconsistent findings observed across our data sets. Future research should conduct systematic, the-

ory-driven investigations into the effect of specific emotion item characteristics on NED results.

Another methodological difference across the present studies concerns the reference timeframe of the AA items. For example, negative emotions and stress were assessed with reference to the previous prompt (Study 1), the entire day (Study 2), or the past hour (Study 3). Retrospective assessments that rely on broader timeframes may increase participants' reliance on generalizations, such as situation-specific or identity-related beliefs, rather than on direct experiential access to current emotions (Robinson & Clore, 2002). In our studies, stress ratings referring to preceding time periods may have captured such generalizations to some extent, potentially altering the relationship between stress and momentary mood. Similarly, the between-person correlation between NED and depression may have been influenced by differences in the timeframes used to assess emotional experiences across the studies.

The studies also differed in the proportion of occasions at which negative emotion were experienced. With 22%, the proportion was especially low in Study 1. Although this may be a realistic proportion of occasions with report of negative emotions (e.g., Zelenski & Larsen, 2000), it may have been affected by our filtering, which only inquired participants to rate their negative emotion intensity if they indicated on a filter item that they had experienced at least one negative emotion, instead of rating them at every occasion. To avoid reinforcing the negation on the filter item (i.e., taking a shortcut within the prompts), we used filler items about the participants' activities. Still, the proportion of occasions with negative emotions remained low. For occasions without reported negative emotions, we set the emotion intensity ratings to zero to indicate that the emotions were experienced "*not at all*". This approach may have biased the NED indices in universes where all occasions were included in the calculation of the ICCs, because it is possible that participants would have responded differently to certain emotion terms if they had encountered them. In universes where exclusively occasions with negative emotions were included, the NED indices might be limited in their validity. The few measurement occasions might not be fully representative of NED, as there are fewer or no occasions at which participants rated all emotions consistently (i.e., zero ratings here).

Although these methodological differences between the studies were not directly incorporated into the multiverse analysis, the inclusion of multiple data sets with varying design features aligns with the concept of a "multiverse of methods" (Harder, 2020). Unlike a classical multiverse analysis, which systematically examines analytic decision combinations within a single data set, a multiverse of methods extends this framework to include design-level variability by analyzing multiple data sets. Our comparison across three studies offers initial insight into how features of study planning and design may affect the robustness of observed effects. Future studies would benefit from a more systematic and comprehensive application of this approach to further investigate the role of methodological variability.

In addition, our study faces specific challenges related to statistical power. Excluding participants for reasons such as low compliance or a negative ICC reduces statistical power, and this limitation is evident in our main analyses. While this reflects typical research practice, it complicates the interpretation of non-significant findings because the compliance threshold decision and power reduction can be confounded. To address this, we artificially increased the Level 2 sample size in exploratory analyses to isolate the effects of data preparation from power issues; the results remained consistent. However, this approach has its own limitations: Resampling participants to increase power can bias the results, since some participants are necessarily drawn multiple times. Thus, the resampled participants are not independent and some participants are weighted more strongly. This may have led to significant results in unexpected directions. We therefore recommend a cautious interpretation of the resampled analyses. Future research should ensure that the power remains adequate in the universe with the strictest compliance threshold (i.e., with the smallest sample size).

Furthermore, in our a priori power analysis, we used an expected effect derived from existing data (also used in Study 2) based on a specific data preprocessing approach. Lafit et al. (2025) recently examined the uncertainty of sample size recommendations based on data from previous AA studies. They demonstrated that factors such as study duration, construct operationalization, and preprocessing choices can affect effect size estimates and their uncertainty (i.e., their confidence intervals), ultimately influencing sample size recommendations. This was also evident in our results: Effect sizes and their confidence intervals varied across data preparation decisions in the multiverse analyses. As a result, different preprocessing choices in the power analysis could have altered the estimated required sample size, raising the possibility that the actual sample size was insufficient in some analytical universes—potentially contributing to non-significant results. Future studies conducting an a priori power analysis based on existing data should explicitly consider the uncertainty in effect estimates introduced by data preparation decisions (for practical guidance, see Lafit et al., 2025).

Constraints on Generality

The generalizability of our findings may be limited for at least two reasons:

First, all three studies utilized non-clinical German samples, which were predominantly female and relatively young, despite efforts to recruit a more diverse group in Study 1 (e.g., by not initially recruiting students) and a more age-heterogeneous sample in Study 2 (by not inviting participants of a certain age to further participation). In general, research suggests gender differences in self-reported emotions (for an overview, see Brody & Hall, 2008) and cross-cultural differences in emotion differentiation (Grossmann et al., 2016). Thus, it is uncertain whether our findings apply to more representative German samples or other cultures, especially given the inconsistent patterns across the relatively similar samples in the present re-

search. Moreover, the current data do not allow conclusions about potential differences in NED between healthy individuals and clinical populations, such as individuals with major depressive disorder, which have been investigated in previous research (e.g., Demiralp et al., 2012).

Second, the results from our multiverse analyses may not be generalizable to other analyses involving mean differences between groups, as our focus was on relationships between NED and adaptive outcomes. ICCs measuring absolute agreement are influenced by the absolute values of emotion intensity ratings, often resulting in smaller values than consistency ICCs, which capture only the covariation between emotion ratings over time. Furthermore, single measurement ICCs tend to be smaller than those based on the average of *k* measurements (Koo & Li, 2016). Taking the absolute value of ICCs (not) into account may be less relevant in relational analyses such as in the present research. It may, however, have a dissimilar impact on between-group differences by affecting distribution and mean levels of ICCs, which might amplify or attenuate such differences.

Implications and Future Directions

The non-robust results across methodological decisions highlight the need for consistent data preparation to ensure the comparability of empirical findings on NED across studies. In accordance with this, the number of arbitrary decisions in a multiverse needs to be reduced to address variation in the results (i.e., the multiverse needs to be *deflated*; Steegen et al., 2016). Potential solutions include refining theory and improving measurement, which could help to identify a superior data preparation method (Steegen et al., 2016). Indeed, consistency ICCs have been recommended over absolute agreement ICCs for NED for theoretical reasons (Erbas et al., 2014; Ottenstein & Lischetzke, 2020), as only the correlations between emotion ratings, rather than their absolute value, are deemed relevant (Feldman Barrett et al., 2001).

Additionally, further research is needed to evaluate the validity of negative ICCs in the context of NED. For instance, future studies could compare participants with negative ICCs to those with positive ICCs close to zero to assess whether their response patterns reflect similarly high levels of differentiation (e.g., Thompson, Liu, et al., 2021). If this interpretation holds, it would be important to examine whether these groups differ systematically on NED-related constructs. This could clarify whether negative ICCs reflect meaningful differences in NED, potentially indicating even higher NED than near-zero positive values. In this case, setting negative ICCs to zero or excluding them could obscure substantive variance and thus be inappropriate. Eliminating such options would also reduce the number of decisions, thereby deflating (i.e., simplifying) the multiverse.

Relatedly, more research is needed to determine how many measurement occasions are required to reliably capture trait-level NED (Thompson, Springstein, et al., 2021). When based on only a few occasions, NED estimates may reflect momentary states rather than a stable trait. Identifying the sufficient number of occasions could help guide future decisions regarding compliance thresholds.

Finally, multiverse analyses using simulated data could help identify which methodological decisions (or combinations there-of) most accurately recover true effects.

Conclusion and Recommendations

Our multiverse analyses of three studies demonstrated that the results of the analyses of interest on NED's adaptive value were largely non-robust across typical data preparation decisions, and there was no discernible, systematic pattern in how the decisions affected the results. Therefore, we recommend to base data preparation in NED research on thorough theoretical consideration rather than relying on procedures reported in previous studies. This is all the more important, the more participants or measurement occasions are affected (e.g., in terms of the number of negative ICCs). Moreover, the findings support the calls for more transparency in emotion differentiation research (Thompson, Springstein, et al., 2021) and in research in general, as researcher degrees of freedom can lead to deviations in substantive conclusions, even in the same data set. Transparent reporting is crucial for assessing the credibility of findings, and we encourage emotion differentiation researchers to preregister their studies and openly report all data preparation choices. Although preregistration reduces decision flexibility in advance, it does not fully eliminate analytic uncertainty, as decisions may still be arbitrary (Hall et al., 2022; Steegen et al., 2016). Therefore, robustness checks, such as multiverse analyses, should be considered whenever no superior data preparation choice is evident.

Author Contribution Statement (CRediT)

Sabrina Ecker: conceptualization, data curation, formal analysis, investigation, project administration, visualization, writing—original draft preparation, writing—review & editing; Charlotte Ottenstein: data curation, investigation, project administration, writing—review & editing; Dominik Vollbracht: investigation, project administration, writing—review & editing; Tanja Lischetzke: conceptualization, funding acquisition, resources, supervision, writing—review & editing

Acknowledgements

We would like to thank Marcel Schmitt for providing the data used in Study 3.

Funding Information

The contributions of Sabrina Ecker and Dominik Vollbracht were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 – Projektnummer 310365261.

Ethics Statement

Study 1 was approved by the local ethics committee of the Department of Psychology at the RPTU University Kaiserslautern-Landau (former University of Koblenz-Landau; approval number LEK-439). In both Study 2 and Study 3, participants consented to the re-use of their data beyond the original study. For details on the ethics approval, see the original articles (Lischetzke et al., 2021; Schmitt et al., 2024).

Competing Interests Statement

The authors have no competing interests to declare.

Data Accessibility

Study 1's design, hypotheses, and data analysis plan were preregistered before data collection started. Studies 2 and 3 were re-analyses of existing data. We have made the preregistration, study materials (codebooks), data sets, data analysis code, and the results of the multiverse analyses accessible on the Open Science Framework (OSF, https://osf.io/jdesw/).

Editors: Wolf Vanpaemel (Senior Editor)

Submitted: November 23, 2024 PDT. Accepted: March 06, 2026 PDT. Published: April 20, 2026 PDT.

References

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

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. The Guilford Press.

Brody, L. R., & Hall, J. A. (2008). Gender and emotion in context. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 395–408). Guilford Press.

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Feldman Barrett, L., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., Deldin, P. J., Gotlib, I. H., & Jonides, J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychological Science*, *23*(11), 1410–1416. https://doi.org/10.1177/0956797612444903

Eckland, N. S., Sperry, S. H., Castro, A. A., & Berenbaum, H. (2022). Intensity, frequency, and differentiation of discrete emotion categories in daily life and their associations with depression, worry, and rumination. *Emotion*, *22*(2), 305–317. https://doi.org/10.1037/emo0001038

Emery, N. N., Simons, J. S., Clarke, C. J., & Gaher, R. M. (2014). Emotion differentiation and alcohol-related problems: The mediating role of urgency. *Addictive Behaviors*, *39*(10), 1459–1463. https://doi.org/10.1016/j.addbeh.2014.05.004

Erbas, Y., Ceulemans, E., Blanke, E. S., Sels, L., Fischer, A., & Kuppens, P. (2019). Emotion differentiation dissected: Between-category, within-category, and integral emotion differentiation, and their relation to well-being. *Cognition & Emotion*, *33*(2), 258–271. https://doi.org/10.1080/02699931.2018.1465894

Erbas, Y., Ceulemans, E., Kalokerinos, E. K., Houben, M., Koval, P., Pe, M. L., & Kuppens, P. (2018). Why I don't always know what I'm feeling: The role of stress in within-person fluctuations in emotion differentiation. *Journal of Personality and Social Psychology*, *115*(2), 179–191. https://doi.org/10.1037/pspa0000126

Erbas, Y., Ceulemans, E., Lee Pe, M., Koval, P., & Kuppens, P. (2014). Negative emotion differentiation: Its personality and well-being correlates and a comparison of different assessment methods. *Cognition & Emotion*, *28*(7), 1196–1213. https://doi.org/10.1080/02699931.2013.875890

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feldman Barrett, L., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion*, *15*(6), 713–724. https://doi.org/10.1080/02699930143000239

Field, A. P. (2005). Intraclass correlation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 948–954). Wiley. https://doi.org/10.1002/0470013192.bsa313

Frijda, N. H. (1993). Moods, emotion episodes, and emotions. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 381–403). Guilford Press.

Giraudeau, B. (1996). Negative values of the intraclass correlation coefficient are not theoretically possible. *Journal of Clinical Epidemiology*, *49*(10), 1205–1206. https://doi.org/10.1016/0895-4356(96)00053-4

Gräfe, K., Zipfel, S., Herzog, W., & Löwe, B. (2004). Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)" [Screening for mental disorders using the "Patient Health Questionnaire (PHQ-D)"]. *Diagnostica*, *50*(4), 171–181. https://doi.org/10.1026/0012-1924.50.4.171

Grossmann, I., Huynh, A. C., & Ellsworth, P. C. (2016). Emotional complexity: Clarifying definitions and cultural correlates. *Journal of Personality and Social Psychology*, *111*(6), 895–916. https://doi.org/10.1037/pspp0000084

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., & Kay, M. (2022). A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum*, *41*(1), 402–426. https://doi.org/10.1111/cgf.14443

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177. https://doi.org/10.1177/1745691620917678

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2022). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*(4), 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

Hoemann, K., Lee, Y., Kuppens, P., Gendron, M., & Boyd, R. L. (2023). Emotional granularity is associated with daily experiential diversity. *Affective Science*, *4*(2), 291–306. https://doi.org/10.1007/s42761-023-00185-2

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863–879. https://doi.org/10.1177/0956797619838763

Kashdan, T. B., Feldman Barrett, L., & McKnight, P. E. (2015). Unpacking emotion differentiation. *Current Directions in Psychological Science*, *24*(1), 10–16. https://doi.org/10.1177/0963721414550708

Klaperski, S., Dawans, B. von, Heinrichs, M., & Fuchs, R. (2013). Does the level of physical exercise affect physiological and psychological responses to psychosocial stress in women? *Psychology of Sport and Exercise*, *14*(2), 266–274. https://doi.org/10.1016/j.psychsport.2012.11.003

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lafit, G., Revol, J., Cloos, L., Kuppens, P., & Ceulemans, E. (2025). The effect of different construct operationalizations, study duration, and preprocessing choices on power-based sample size recommendations in intensive longitudinal research. *Assessment*, *32*(2), 206–223. https://doi.org/10.1177/10731911241286868

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

Lazarus, G., & Fisher, A. J. (2021). Negative emotion differentiation predicts psychotherapy outcome: Preliminary findings. *Frontiers in Psychology*, *12*, 689407. https://doi.org/10.3389/fpsyg.2021.689407

Lennarz, H. K., Lichtwarck-Aschoff, A., Timmerman, M. E., & Granic, I. (2018). Emotion differentiation and its relation with emotional well-being in adolescents. *Cognition & Emotion*, *32*(3), 651–657. https://doi.org/10.1080/02699931.2017.1338177

LimeSurvey GmbH. (2022). *LimeSurvey: An open source survey tool* (5.3.24) [Computer software]. Limesurvey GmbH. http://www.limesurvey.org

Lischetzke, T., & Könen, T. (2022). Mood. In F. Maggino (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1–6). Springer. https://doi.org/10.1007/978-3-319-69909-7_1842-2

Lischetzke, T., Schemer, L., Glombiewski, J. A., In-Albon, T., Karbach, J., & Könen, T. (2021). Negative emotion differentiation attenuates the within-person indirect effect of daily stress on nightly sleep quality through calmness. *Frontiers in Psychology*, *12*, 684117. https://doi.org/10.3389/fpsyg.2021.684117

Liu, D. Y., Gilbert, K. E., & Thompson, R. J. (2020). Emotion differentiation moderates the effects of rumination on depression: A longitudinal study. *Emotion*, *20*(7), 1234–1243. https://doi.org/10.1037/emo0000627

Matt, L. M., Fresco, D. M., & Coifman, K. G. (2016). Trait anxiety and attenuated negative affect differentiation: A vulnerability factor to consider? *Anxiety, Stress, and Coping*, *29*(6), 685–698. https://doi.org/10.1080/10615806.2016.1163544

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Mikhail, M. E., Keel, P. K., Burt, S. A., Neale, M., Boker, S., & Klump, K. L. (2020). Low emotion differentiation: An affective correlate of binge eating? *The International Journal of Eating Disorders*, *53*(3), 412–421. https://doi.org/10.1002/eat.23207

Nilges, P., & Essau, C. (2021). *DASS. Depressions-Angst-Stress-Skalen - deutschsprachige Kurzfassung [DASS. Depression-anxiety-stress scales - German short version]*. Open Test Archive. https://doi.org/10.23668/psycharchives.4579

Nook, E. C., Flournoy, J. C., Rodman, A. M., Mair, P., & McLaughlin, K. A. (2021). High emotion differentiation buffers against internalizing symptoms following exposure to stressful life events in adolescence: An intensive longitudinal study. *Clinical Psychological Science*, *9*(4), 699–718. https://doi.org/10.1177/2167702620979786

Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2018). The nonlinear development of emotion differentiation: Granular emotional experience is low in adolescence. *Psychological Science*, *29*(8), 1346–1357. https://doi.org/10.1177/0956797618773357

O'Brien, S. T., Dozo, N., Hinton, J. D. X., Moeck, E. K., Susanto, R., Jayaputera, G. T., Sinnott, R. O., Vu, D., Alvarez-Jimenez, M., Gleeson, J., & Koval, P. (2024). Sema3: A free smartphone platform for daily life surveys. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-024-02445-w

O'Toole, M. S., Renna, M. E., Elkjær, E., Mikkelsen, M. B., & Mennin, D. S. (2020). A systematic review and meta-analysis of the association between complexity of emotion experience and behavioral adaptation. *Emotion Review*, *12*(1), 23–38. https://doi.org/10.1177/1754073919876019

Ottenstein, C., & Lischetzke, T. (2020). Development of a novel method of emotion differentiation that uses open-ended descriptions of momentary affective states. *Assessment*, *27*(8), 1928–1945. https://doi.org/10.1177/1073191119839138

Ottenstein, C., & Werner, L. (2022). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

R Core Team. (2025). *R: A language and environment for statistical computing* (4.5.0) [Computer software]. https://www.R-project.org/

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research* (2.2.9). Computer software. https://CRAN.R-project.org/package=psych

Rights, J. D., & Sterba, S. K. (2020). New recommendations on the use of R-squared differences in multilevel model comparisons. *Multivariate Behavioral Research*, *55*(4), 568–599. https://doi.org/10.1080/00273171.2019.1660605

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*(6), 934–960. https://doi.org/10.1037//0033-2909.128.6.934

Sacharin, V., Schlegel, K., & Scherer, K. R. (2012). *Geneva emotion wheel rating study*. University of Geneva, Swiss Center for Affective Sciences. http://archive-ouverte.unige.ch/unige:97849

Schmitt, M. C., Vogelsmeier, L. V. D. E., Erbas, Y., Stuber, S., & Lischetzke, T. (2024). Exploring within-person variability in qualitative negative and positive emotional granularity by means of latent markov factor analysis. *Multivariate Behavioral Research*, *59*(4), 781–800. https://doi.org/10.1080/00273171.2024.2328381

Schreuder, M. J., Wichers, M., Hartman, C. A., Menne-Lothmann, C., Decoster, J., van Winkel, R., Delespaul, P., Hert, M. de, Derom, C., Thiery, E., Rutten, B. P. F., Jacobs, N., & van Os, J. (2022). Lower emotional complexity as a prospective predictor of psychopathology in adolescents from the general population. *Emotion*, *22*(5), 836–843. https://doi.org/10.1037/emo0000778

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 289–308). Sage. https://doi.org/10.4135/9781446249215.n15

Seah, T. H. S., & Coifman, K. G. (2022). Emotion differentiation and behavioral dysregulation in clinical and nonclinical samples: A meta-analysis. *Emotion*, *22*(7), 1686–1697. https://doi.org/10.1037/emo0000968

Shaw, M., Rights, J. D., Sterba, S. S., & Flake, J. K. (2023). R2mlm: An R package calculating R-squared measures for multilevel models. *Behavior Research Methods*, *55*(4), 1942–1964. https://doi.org/10.3758/s13428-022-01841-4

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA*, *282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Starr, L. R., Hershenberg, R., Li, Y. I., & Shaw, Z. A. (2017). When feelings lack precision: Low positive and negative emotion differentiation and depressive symptoms in daily life. *Clinical Psychological Science*, *5*(4), 613–631. https://doi.org/10.1177/2167702617694657

Starr, L. R., Hershenberg, R., Shaw, Z. A., Li, Y. I., & Santee, A. C. (2020). The perils of murky emotions: Emotion differentiation moderates the prospective relationship between naturalistic stress exposure and adolescent depression. *Emotion*, *20*(6), 927–938. https://doi.org/10.1037/emo0000630

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF): Handanweisung [The multidimensional mood state questionnaire (MDBF): Manual]*. Hogrefe.

Thompson, R. J., Liu, D. Y., Sudit, E., & Boden, M. (2021). Emotion differentiation in current and remitted major depressive disorder. *Frontiers in Psychology*, *12*, 685851. https://doi.org/10.3389/fpsyg.2021.685851

Thompson, R. J., Springstein, T., & Boden, M. (2021). Gaining clarity about emotion differentiation. *Social and Personality Psychology Compass*, *15*(3). https://doi.org/10.1111/spc3.12584

Vollbracht, D., Ottenstein, C., Ecker, S., & Lischetzke, T. (2026). Slider vs. Likert scales: Psychometric properties in ambulatory assessment. *PsyArXiv*. https://doi.org/10.31234/osf.io/y29xa_v1

Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63–83. https://doi.org/10.1037/met0000030

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01777-1

Williams, G. E., & Uliaszek, A. A. (2022). Measuring negative emotion differentiation via coded descriptions of emotional experience. *Assessment*, *29*(6), 1144–1157. https://doi.org/10.1177/10731911211003949

Willroth, E. C., Flett, J. A. M., & Mauss, I. B. (2020). Depressive symptoms and deficits in stress-reactive negative, positive, and within-emotion-category differentiation: A daily diary study. *Journal of Personality*, *88*(2), 174–184. https://doi.org/10.1111/jopy.12475

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

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