VAS vs. Likert in ESM Study: The Effect of the Choice of Response Format on Extreme Response Style

Faculty of Behavioural, Management, and Social Sciences

Department of Psychology, University of Twente

202000384: BSc Thesis Psychology

Nick A. Delventhal

BSc. Psychology

Supervisor: Dr. Jannis T. Kraiss

Second supervisor: Dr. Thomas R. Vaessen

June 27, 2023

Abstract

Background: The Experience Sampling Method (ESM) is becoming increasingly popular as a granular measure of emotional experience. However, there is a lack of knowledge on the effect of research design on the results. Specifically, it is unknown how choosing a Visual Analogue Scale (VAS) versus a Likert scale influences Extreme Response Style (ERS). Therefore, this research aims to investigate the effect of using a VAS compared to a Likert scale on participants' proportion of extreme responses in ESM studies.

Methods: Participants were recruited in three cohorts through convenience and voluntary sampling, resulting in a sample of 89 participants (Likert $n = 53$, VAS $n = 36$, mean age 29.94, 63.49% female, 36.51% male). This study designed a randomised controlled trial with two conditions (VAS and Likert) with a between-subject design. Generalised linear mixed models (GLMM), both for the overall effect and at the individual item level, examined the effect of using a VAS versus a Likert scale on the proportion of extreme responses.

Results: Responses in both conditions were not normally distributed and strongly zero-inflated. The items showed low (.291) to very high (.956) inter-item correlations. The GLMM results showed no significant effect of using a VAS versus a Likert scale (b = -.191, p = .756), the effect was also not significant across individual items ($b = -.556$ to $b = -.114$, $p = .229$ to $p = .771$.).

Conclusion: These findings tentatively suggest that choosing a VAS versus Likert scale does not have an impact on ERS in ESM research. Nonetheless, the non-representative sample and potential exclusion bias, warrant caution in generalizing the findings.

VAS vs. Likert in ESM Study: The Effect of Choice of Response Format on Extreme Response Style

"On a scale from one to seven, please rate how you felt the past two weeks". This is obviously impossible to answer for most people in most situations. But even if asked how people felt just 12 hours ago, they have trouble giving an accurate answer. Many participants simply cannot remember and even if they do, their memory can be subject to systematic tendencies that distort the information in hindsight (Beal, 2015; Christensen-Szalanski & Willham, 1991; Klein, 2015).

For giving researchers a reliable insight into the momentary experiences and emotions of individuals in their daily life, the Experience Sampling Method (ESM), also called ecological momentary assessment, has increased in popularity in the recent century (van Berkel et al., 2018). ESM involves repeatedly prompting participants to rate or express their current thoughts and feelings in the context of their daily routine (Myin-Germeys & Kuppens, 2022). Compared to retrospective collection methods, ESM has the advantage of reducing memory bias and thereby increasing ecological validity. (Csikszentmihalyi, 2014; Myin-Germeys & Kuppens, 2022; Versluis et al., 2021).

To be able to quantify the subjective experiences of participants, researchers can use a variety of response scales, the most common being a Likert scale and Visual Analogue Scale (VAS). The Likert scale is a discrete scale that is one of the most established psychometric measurements in psychology and is easy to use for participants and data analysts alike (Guyatt et al., 1987). Most often, a 7-point Likert scale is used on which participants can choose between seven response options, representing s*trongly disagree* to *strongly agree.* The VAS often spans from 0 to 100, however, participants cannot see the 101 response options, rather they see a horizontal continuous line with labelled anchors indicating the extreme points (e.g., *not at all* to *very much*). On this line, participants position, or slide, the indicator to the location that they consider representative of the intensity of their experience. (Gries et al., 2018).

When comparing the two response scales in retrospective studies, *reliability*, *validity*, *means*, and *standard deviation* were found to be analogous by most researchers (Guyatt et al., 1987; Lukacz et al., 2004). However, the Likert scale is often the preferred option because of its simplicity and reputation, while VAS is increasing in popularity mostly due to its sensitivity (Kuhlmann et al., 2017).

Nonetheless, research within the context of ESM that pertains to a comparison of these two measurements is still largely missing (Eisele et al., 2022). Hence, researchers do

not have the possibility to inform themselves about the effects of response scales on participants as there is not a sufficient body of research comparing different measures for ESM studies (Janssens et al., 2020; Myin-Germeys & Kuppens, 2022).

This is problematic as it is suspected that the choice of response scale has an effect on the distribution of the data in ESM studies (Eisele et al., 2022). One possible avenue of influence is that participants respond more often at the endpoints of the scale when using one scale, compared to another. The underlying response behaviour is called Extreme Response Style (ERS). ERS is the bias of participants to gravitate towards the endpoints on a response scale, more than their true belief, attitude, or behaviour would indicate, which results in them choosing the extreme response options more frequently than the middle options (Cabooter et al., 2016; Kieruj & Moors, 2013).

While there is no research on the effect of choice of response scale on ERS, there are indications, based on findings in retrospective studies, that suggest that using a VAS compared to a Likert scale could bias the results through an increased rate of extreme responses.

Firstly, the anchors used in a VAS might impact participants' responses. The use of anchors only at the extreme points, such as *not at all* and *very much*, may encourage participants to use extreme values more often compared to when responding to a Likert scale which uses seven visual spatial anchors that also correspond to "somewhat" and "moderately" (Bishop & Herron, 2015; Couclelis et al., 1987). The visible difference in response options could encourage participants to use an ERS more frequently when using a VAS.

Secondly, further amplifying this potential effect of response scales on ERS, Funke & Reips (2012) found that adequately responding to a VAS requires more cognitive effort than a Likert scale due to the larger number of response options available. Therefore, answering the VAS is a process less guided and thus more difficult. This can lead participants with low motivation or concentration to become frustrated to the point that they might use a potentially biased response heuristic (Kieruj & Moors, 2013). Krosnick (1991), coined the term *satisficing* to describe this phenomenon, where participants who lack the willingness or time to put in the necessary effort to form optimal answers to attitude questions, might use heuristic shortcuts to create answers that are not necessarily reflective of their actual experience, belief, or attitude. One such heuristic shortcut is ERS, thus, when participants already have a tendency to an ERS the occurrence of it might be more frequent for the more difficult VAS (Kieruj & Moors, 2013).

Additionally, it is worth noting that ERS is per definition independent of the content of the question, as the style of responding is determining the response and not the content itself (Myin-Germeys & Kuppens, 2022). Therefore, it is recommended to measure ERS on a wide range of topics with questions that are not at all or only slightly related (Greenleaf, 1992). However, an ESM study is asking about the immediate experiences of individuals which makes the construction of uncorrelated items difficult to find in existing literature (Dejonckheere & Erbas, 2022). Nonetheless, it is important to recognise and deal with this potential inaccuracy of the measurement of ERS. Therefore, each item should be investigated individually to observe if ERS is found consistently across the ESM questions, to account for any single question content effect. Also generally, the composite of all items may mask the variability and nuances of the often complex and heterogeneous ESM data (McNeish & Wolf, 2020). Hence, the ESM data should be investigated not just in terms of the overall trend but also on the basis of each item.

To summarise, it was identified that scientific knowledge of the response scale options for ESM studies is lacking. Specifically, research comparing Likert scale and VAS in their effects on participants' ERS is non-existent, even in retrospective studies. This is problematic as the choice of response format becomes a matter of personal preference which neglects the potential implications of different scales.

To investigate this, this research will ask: "Does using a VAS compared to a Likert scale in an ESM study lead to a greater proportion of extreme responses for positive and negative affect and stress?". To answer this question, the following hypothesis will be tested: "Using a VAS in an ESM study will lead to significantly greater proportion of extreme responses when compared to Likert scale.".

In addition to investigating the overall distribution, this research will also examine the effect of response format on each of the items individually. Therefore, it will be asked: "What is the effect of response format on each of the nine items individually?". To answer this question, the following hypothesis will be tested: "Each of the 9 items in the ESM study will individually show a significantly greater proportion of extreme responses when using a VAS compared to a Likert scale.".

Methods

Design

Ethical approval was granted by the Ethics Committee of Behavioral, Management and Social Sciences of the University of Twente (request number: 221244).

To test the effect of a Likert scale compared to a VAS on ERS in the context of the ESM, an ESM study was set up with a between-subject design consisting of two conditions (Likert scale vs. VAS).

Participants

To achieve the power of .8 with the effect size of f = .25, it was aimed to have at least a total of 128 retained participants for the analysis. To account for participants that might drop out or only give insufficient answers, the recruitment minimum target was set to 136 participants, as Vachon et al. (2019) found that on average ESM studies reported retention rates of around 94%. Participants were recruited from the social environment of the researchers or via the SONA system of the University of Twente. On there, participants would obtain 1.5 points that count towards the mandatory balance of at least 15 which all bachelor students will have needed to obtain at the end of their bachelor. To incentivise partaking in the study, all participants were entered into a lottery for a 40€ Amazon or Bol voucher to win. Inclusion criteria were a minimum age of 18 years, possession of a smartphone, and English proficiency.

Randomization

This research used stratified randomisation by creating a randomisation cluster for each of the researcher's groups of participants. These clusters of participants were then randomised evenly into each condition (Sealed Envelope Ltd., n.d.). This accounts for the potential bias that could arise from selecting participants who share similarities within the groups of individuals that were recruited from each researcher.

Materials

The baseline questionnaire administered prior to the ESM study gathered demographic information and assessed mental well-being through several measures. The ESM questionnaires inquired about the participants' current emotional state, present company, the most recent significant event, and their coping mechanisms. However, only the items measuring momentary affect were deemed relevant to the study.

In line with previous ESM studies, momentary affect was operationalised as positive and negative affect and stress (M. Janssens et al., 2020; van der Linden et al., 2021). To measure affect the items from the Positive and Negative Affect Schedule (PANAS) were used (Watson et al., 1988). The positive affect item consisted of *"How cheerful / enthusiastic / satisfied / relaxed do you feel right now?"*. The negative affect items consisted of: "*How anxious / irritable / down / guilty / do you feel right now?*". Stress was measured using self-perceived level of stress asking: *How stressed do you feel right now?*".

For participants in the VAS and Likert condition, the questions were administered with a 7-point Likert scale and on a VAS ranging from 0 to 100, respectively. Both scales included labels only at the endpoints (*not at all, very much*).

Procedure

Before the start of the study, participants received an E-Mail with the most important information about the study and access to the app Ethica (see Appendix A & B; Ethica Data Services Inc., 2023). In the app, the participants were asked to give their informed consent (see Appendix C) and complete the baseline questionnaire and the ESM questionnaires. The ESM questionnaires asked participants to fill out 10 questionnaires a day for 7 days. This created up to 70 responses per participant and is a commonly used assessment frequency in ESM studies (Dejonckheere & Erbas, 2022). These questionnaires were randomized within 10 equally sized time intervals from 7:30 AM to 10:30 PM, so participants received one questionnaire at a random moment in time within each of the intervals (see Table 1). Randomizing questionnaires within equally sized time intervals allows for representative sampling of participants' experiences and controls their distribution over the whole day.

Table 1

Schedule of the Questionnaires

| Day	Questionnaire	Triggered	Expire
1	Baseline Questionnaire	Once at the beginning of the study	No
Every day (7 days)	ESM Questionnaire	7:30 – 9:00 9:00 – 10:30 12:00 – 13:30 13:30 – 15:00 15:00 – 16:30 16:30 – 18:00 18:00 – 19:30 19:30 – 21:00 21:00 – 22:30	Yes, after 15 minutes

Data collection was conducted in three cohorts between November 2022 and April 2023 (see Figure 1). As only participants from one condition of the first cohort were useful for the analysis, the second cohort was randomized at a 30/70 ratio. However, this did not fully achieve the desired balance, as the VAS condition ended up with more participants (see Figure 1). The current cohort was randomized at a 50/50 ratio.

Data Analysis

Prior to the recruitment of the recent cohort, a power analysis was conducted using the program G*Power (Faul et al., 2009). To obtain a level two between-subject sample size approximation the Statistical test *ANOVA: Fixed effects, omnibus, one-way* was chosen with the effect size of .25 and power of .8.

Figure 1

Participants per Cohort and Condition Before and After Data Cleaning
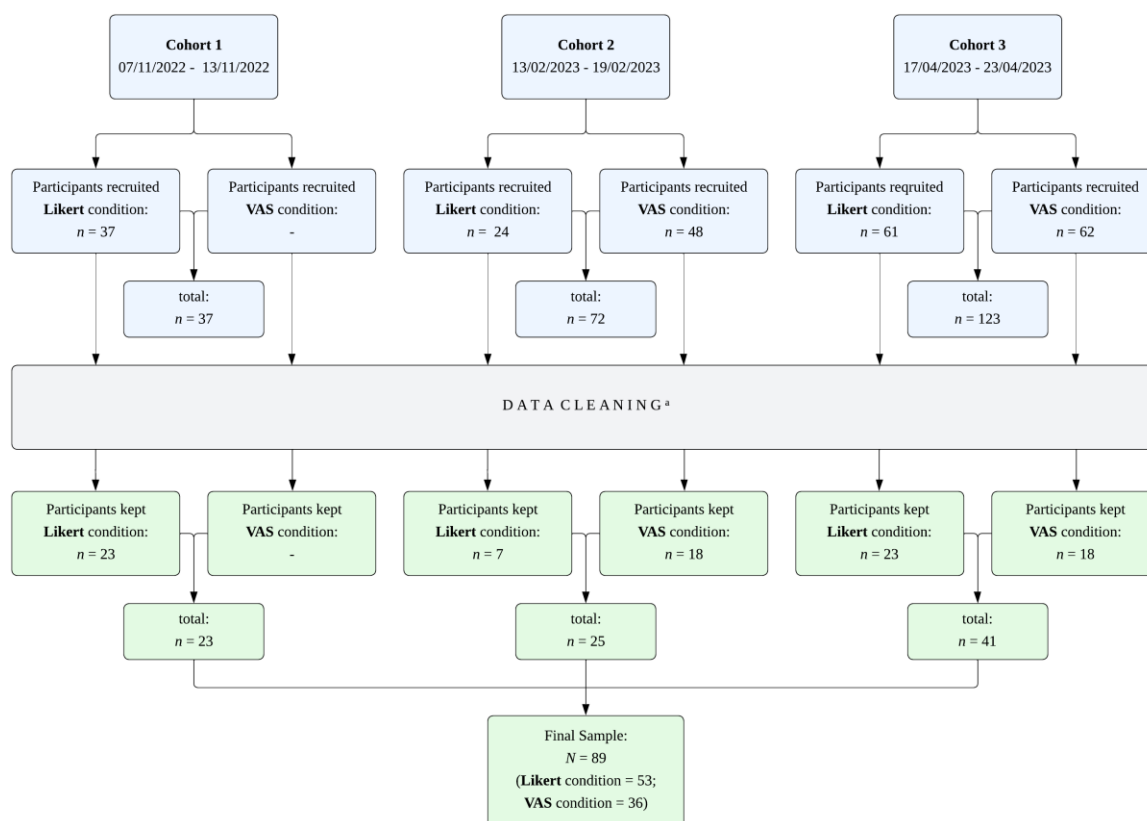
Note. [a] Participants were removed that complete less than 1/3 of assessments.

All data cleaning and analyses were performed in the Statistical Software R (R Core Team, 2018). To clean the data, all responses of individual participants who completed less than 1/3 of the questionnaires were removed from the analysis. Only 1/3 of the answers were deemed to not be representative of the whole duration of the study, which then makes the data inappropriate to compare to other participants' data (Viechtbauer, 2022). The missing observations from retained participants were dealt with in each of the analyses using full case analysis. To test if the dropout rates differed significantly between conditions, a Pearson's Chi-squared test was conducted. Additionally, Yates' continuity correction was also included to account for the potential bias introduced when the sample size is small. To visually receive a first understanding of the data, histograms were created showing the frequency of responses for each item for both conditions. Additionally, the inter-item correlations of all items were calculated, and the average inter-item correlation was retrieved.

For the measurement of ERS, the proportion score of extreme responses is a commonly used measurement (Greenleaf, 1992). For the Likert scale, all responses that are either one or seven were seen as extreme responses. For the VAS, the 101 response options were divided into seven categories and thus, similar to the Likert scale all responses in categories one and seven were defined as extreme responses (category 1 = 0-13, category 7 = 87-100). The proportion was calculated by dividing the number of extreme responses by the total number of responses per assessment.

For the first hypothesis, the proportion of extreme responses was counted per condition (Likert and VAS) and compared in a boxplot. For the main analysis, a generalised linear mixed model (GLMM) using a logistic regression analysis was constructed, using the R package "lme4" (Bates et al., 2015). The fixed effect was condition as an independent binary variable (Likert =1, VAS = 2) and the proportion of extreme responses was the dependent continuous variable. The random effect was participant, to account for the nested data within participants due to the nature of ESM measurements. The GLMM was chosen over a conventional linear mixed model, as it is recommended for proportion scores and is specifically suitable for data that ranges between 0 and 1 that is not normally distributed (Chen et al., 2017). To visualise the effect, a scatterplot with a regression line was constructed, using the R package "ggplot2" (Wickham & Hadley, 2016).

For the second hypothesis, the proportion of extreme responses was added as a binary variable for every single response indicating it to be an extreme response (1) or not (0). This allowed to retrieve the proportion of extreme responses per item. With this, nine GLMMs were constructed. The fixed effect for each GLMM condition was the independent binary

variable (Likert =1, VAS = 2) and the proportion of extreme responses per item was the dependent variable, while the random effect was participant.

Results

A total of 232 participants were recruited. 37 in the first cohort, 72 in the second, and 123 in the third. During data analysis, 143 participants were excluded because they had completed less than 33.33% of the questionnaires. Of the remaining participants, 36 were located in the VAS condition and 53 in the Likert condition (see Figure 1). Thus, 67.27% of the VAS condition and 56.56% in the Likert condition had to be removed. After conducting Pearson's Chi-squared test with Yates' continuity correction, this difference in dropout rate was found to be non-significant, $\chi^2(1) = 2.374$, $p = .123$. Finally, a sample of 89 participants was retained for the analysis. This was 39 participants short of the 128 aimed for participants based on .25 effect size and power of .8.

Because 29 participants did not complete the baseline questionnaire, the demographic characteristics of the sample could only be investigated on the basis of 63 participants. Forty participants were females (63.49%) and 23 males (36.51%), aged between 20 and 67 years (Mean = 29.94, SD = 13.79). The descriptive statistics in regard to each condition can be seen in Table 2.

The average inter-item correlation was found to be high with a value of .611. Specifically, the PA items showed a very high correlation of .916 and the NA items had a high correlation of .710. For the full correlation matrix see Table 3.

While examining the nature of the responses per item, a strong zero inflation can be seen in all negative Affect and stress items (see Figure 2 & 3). The proportion of extreme responses for VAS was found to be .299 and for Likert .336 across all items and responses (see Figure 4).

First Hypothesis

In the main analysis for the first hypothesis, the GLMM showed that using a VAS instead of Likert scale led to a smaller proportion of extreme responses, however, the effect was found to be non-significant ($b = -.191$, $p = .756$). Results of the GLMM can be found in Table 4 and a visualisation in Figure 5, displaying the slightly negative slope of VAS on the proportion of extreme responses.

Table 2

Descriptive Statistics of the Sample that Completed the Baseline Questionnaire

Variable	Likert (*n* = 28)		VAS (*n* = 35)	
Mean Age (*SD*)	29.9 (13.7)		30 (14)	
	n	%	*n*	%
Gender				
Female	17	60.71	23	65.71
Male	11	39.29	12	34.29
Nationality				
German	27	96.43	34	97.14
Dutch	1	3.57	1	2.86
Occupation				
Working	12	42.86	10	28.57
Student	8	28.57	12	34.29
Studying and Working	8	28.57	11	31.43
Self-employed	0	0	1	2.86
Other	0	0	1	2. 86
Education (highest degree completed)				
High School	9	32.14	16	45.71
Bachelor	13	46.43	11	31.43
Master	5	17.86	8	22.86
Other	1	3.57	0	0

Note. N = 63. *N* for the demographic data is lower than for the analysis as not all participants indicated demographic data.
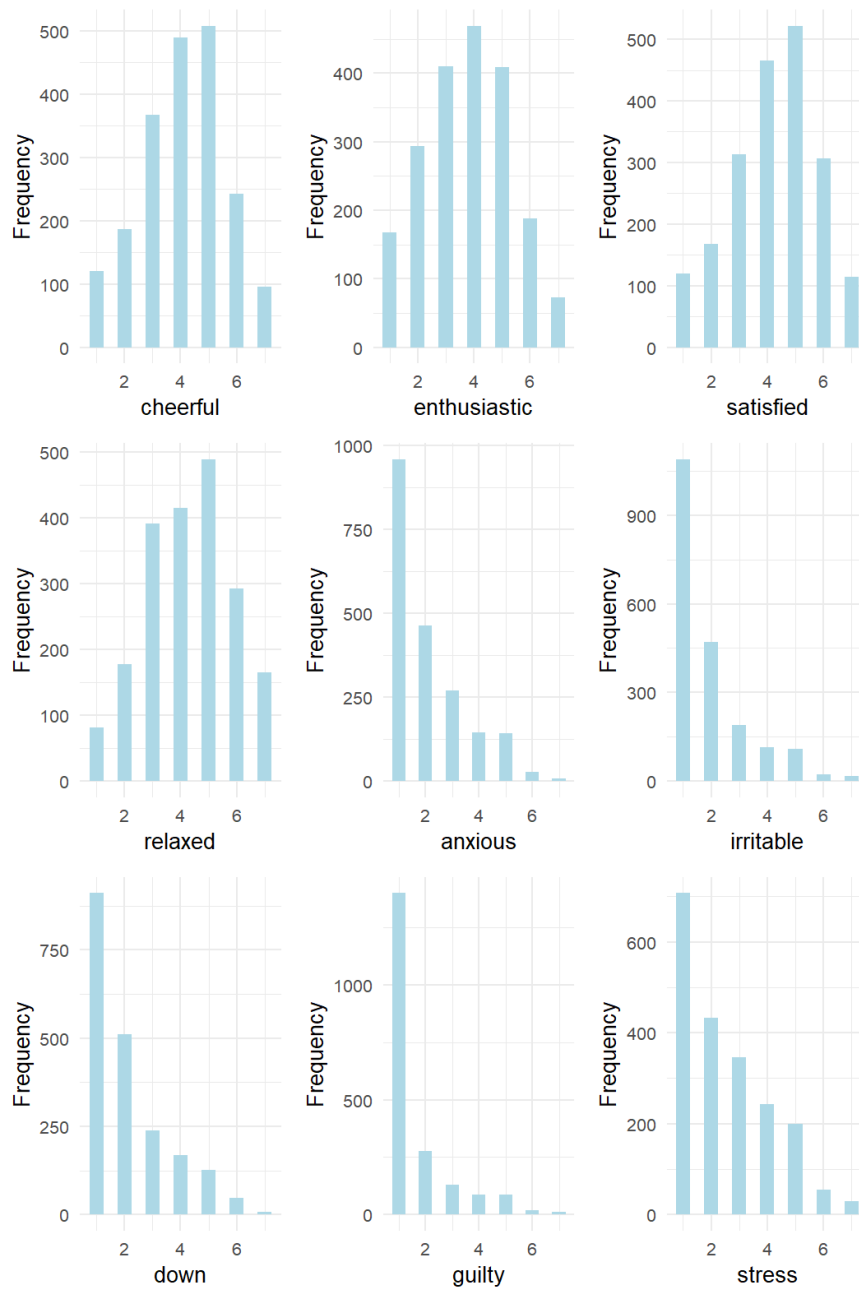
Figure 2

Frequency of Responses for Each Item for the Likert Condition
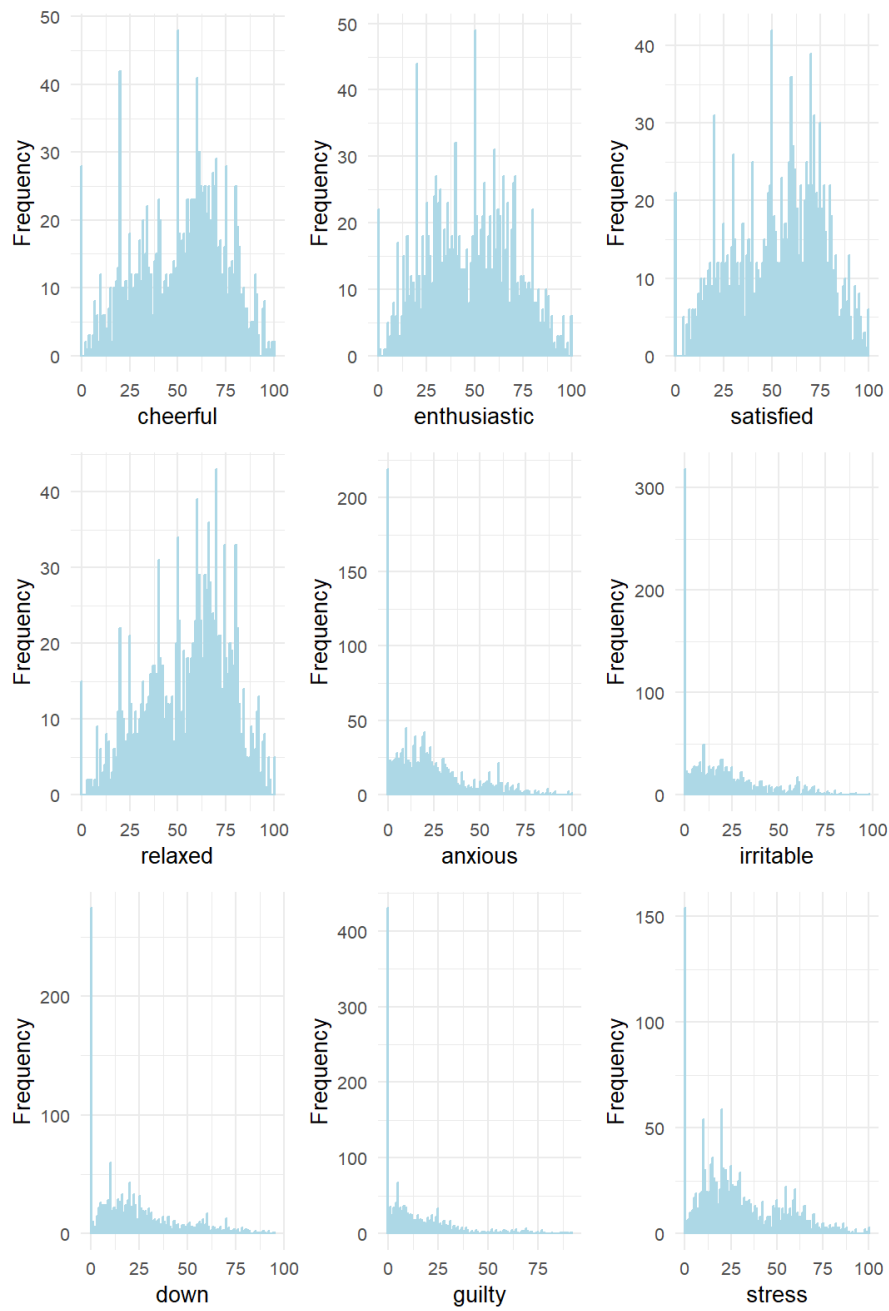

Note. Level 1 $N = 1317$.

Figure 3

Frequency of Responses for Each Item for the VAS Condition

Note. Level 1 *N* = 1964.

Table 3

Inter-Item Correlation Matrix

| Item	cheer-ful	enthu-siastic	satisfied	relaxed	an-xious	irri-table	down	guilty	stress
cheerful	1*								
enthu-siastic	.956*	1*							
satisfied	.938*	.925*	1*						
relaxed	.888*	.867*	.921*	1*					
anxious	.431*	.42*	.404*	.391*	1*				
irritable	.366*	.362*	.348*	.346*	.771*	1*			
down	.37*	.338*	.35*	.38*	.765*	.788*	1*		
guilty	.325*	.313*	.289*	.291*	.644*	.659*	.636*	1*	
stress	.463*	.457*	.428*	.401*	.796*	.785*	.751*	.682*	1*

Note. $N = 3294$. * Significant correlation, $p < .05$.

Second Hypothesis

For the second hypothesis, the GLMMs of all negative affect items, as well as the *stress* and *cheerfulness* items, showed a non-significant effect ($b = -.556$ to $b = -.114$, $p = .229$ and $p = .771$). For the three positive affect items *enthusiastic*, *satisfied*, and *relaxed* problems during the analysis occurred. It was indicated that the models failed to converge and were nearly unidentifiable due to very large Eigenvalues. Nonetheless, extremely low standard deviations were retrieved alongside extremely low p-values ($SE = <.001$, $p = <2e-16$). However, these p-values below the threshold of significance should not be viewed as suggesting a significant effect but rather as a result of the failure of the model. Conversely, the very large eigenvalues suggest very high collinearity of the predictor condition which should rather be seen as an indicator that the effect of condition on proportion of extreme responses is in fact not existent for these items specifically. To support this assumption, a very similar proportion of extreme responses for these items that were retrieved which can be viewed alongside the results of the GLMMs in Table 3

Figure 4

Composite Score of Proportion of Extreme Responses per Condition
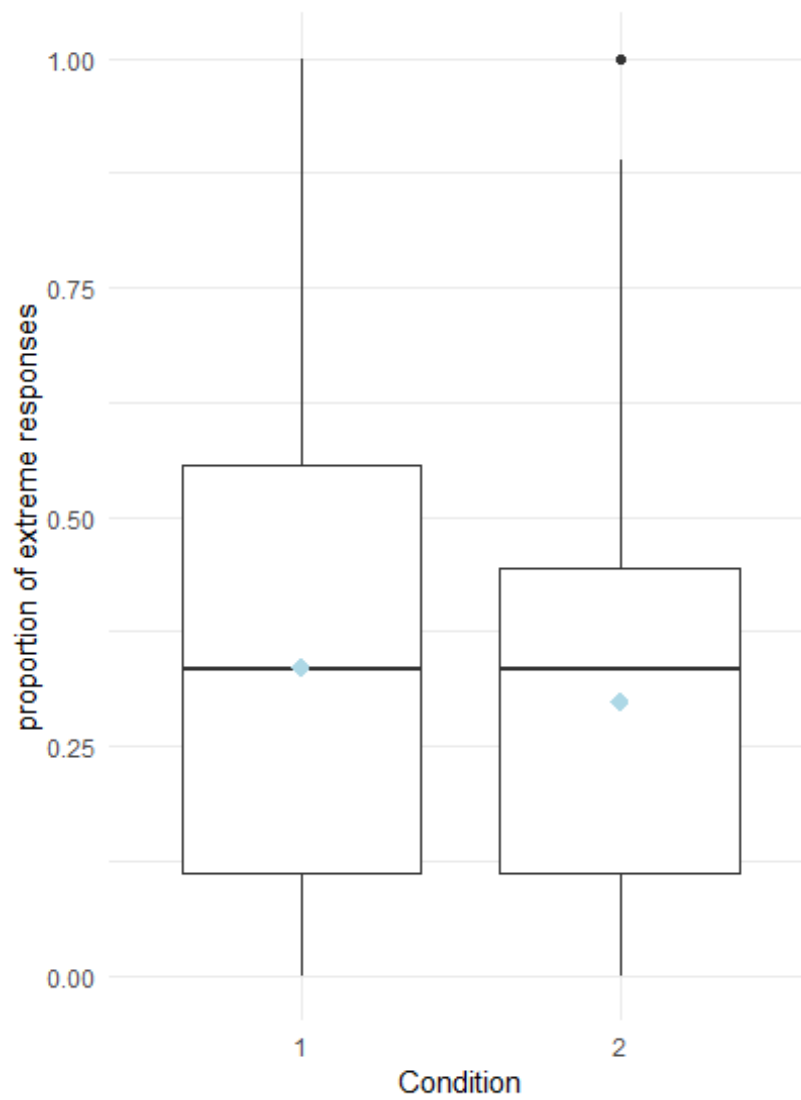
Note. Level 1 $N = 3294$. 1 = Likert condition, 2 = VAS condition. The light blue diamond represents the mean.

Table 4

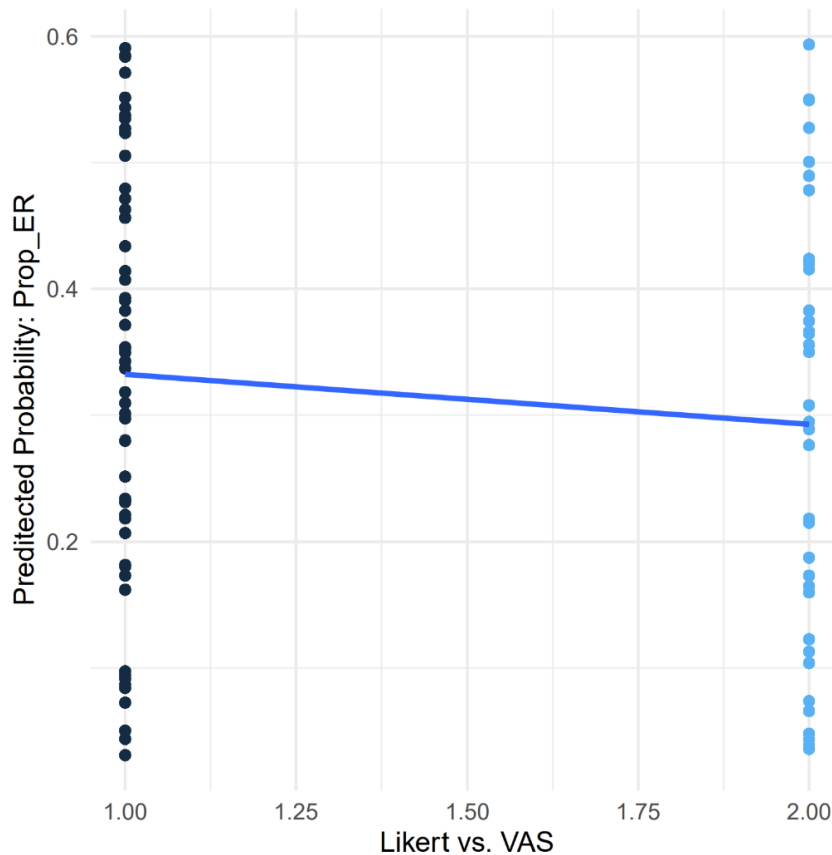*Proportion of Extreme Responses per Condition and GLMMs per Item and Overall*

Item	Proportion of extreme responses		Estimate	SE	df	95% CI		p
	Likert	VAS				LL	UL	
cheerful	.108	.104	-.114	.196	3412	-.898	.666	.771
enthusiastic	.120	.109	-2.866	<.001	3407	-ᵃ	-ᵃ	<2e-16
satisfied	.117	.116	.067	<.001	3402	-ᵃ	-ᵃ	<2e-16
relaxed	.122	.101	-.149	<.001	3396	-ᵃ	-ᵃ	<2e-16
anxious	.480	.409	-.479	.521	3395	-1.527	.562	.357
irritable	.549	.485	-.556	.463	3392	-1.486	. 362	.229
down	.457	.426	-.330	.466	3389	-1.270	. 595	.479
guilty	.702	.635	-.457	.625	3383	-1. 727	.812	.464
stress	.366	.302	-.434	.400	3382	-1.236	.360	.278
Total	. 336	.299	-.191	.615	3378	-1.450	1.130	.756

Note. Level 1 $N = 3294$. CI = confidence interval; LL = lower limit; UL = upper limit.

[a] Due to the underlying data, it was not possible to retrieve LL and UL from these items.

Figure 5

Regression Line GLMM for Proportion of Extreme Responses per Condition

Note. N = 3294. Prop_ER = Proportion of extreme responses.

Discussion

While it was expected that using a VAS instead of a Likert scale in an ESM study would lead to more ERS, this research was unable to observe a significant effect of the response scale on ERS. Moreover, this non-significant effect was replicated across all items individually as well. Therefore, both hypotheses: "Using a VAS in an ESM study will lead to a significantly greater proportion of extreme responses when compared to Likert scale." and "Each of the 9 items in the ESM study will individually show a significantly greater proportion of extreme responses when using a VAS compared to a Likert scale." must be rejected.

Main Findings and Directions for Further Research

These findings are contrary to the expectations put forth by Eisele (2022), who hypothesized that response format would have an effect on the results. The results of this study also contradict the indications found in retrospective studies. For instance, the inclusion

of labels only at the endpoints of the VAS was expected to lead to greater usage of response heuristics, such as ERS. Furthermore, it was hypothesised that the finding that VAS is more cognitively demanding (Funke & Reips, 2012), would constitute another avenue of affecting ERS. However, the underlying assumption put forward by Funke and Reips (2012), could not be observed through any notable difference in the data. While a greater number of participants assigned to the VAS condition completed their assessment with insufficient frequency, this difference in response frequency between the VAS and Likert conditions was found to be non-significant.

Viewing these outcomes in light of the mentioned findings, that generally the psychometric properties of VAS and Likert scale were found to be similar (Guyatt et al., 1987b; Kuhlmann et al., 2017; Lukacz et al., 2004), it could be implied that the frequency of extreme responses is not dependent on the response format but that the variation in the data can be explained by other factors. This would be in line with Kieruj & Moors (2013) who found in a retrospective study that ERS is first and foremost a matter of personal style and not study design.

However, caution should be exercised in suggesting the interchangeability of VAS and Likert scales in the context of ESM. The non-significance could also be explained by the modality of the measurement of ERS. Usually, ERS is measured on a wide range of topics with items correlated as little as possible to account for probable content effects (Greenleaf, 1992). However, this study only relied on items measuring positive and negative affect and stress, from which many were highly correlated. While it was controlled for that no single item overly influenced the overall effect, this measure of ERS could still be subject to effects of the rather narrow topic. Hinting in this direction is the strong zero inflation observed in the negative affect items. This could suggest that participants from both conditions naturally tended towards the endpoint of the scale due to the nature of the content. This might have undermined individuals who would have chosen extreme answers due to the influence of the response format and made this number of participants appear insignificant in comparison. An example of how content effects were avoided can be seen in the research of Laerhoven et al. (2004), who used relatively unrelated questions about simple activities, feelings, and opinions to obtain results that are as unrelated to the content as possible.

Additionally, specific moderating factors, such as personality factors or cultural influences, have been identified in the literature as influential in shaping extreme responses which were not recognised by the group model of the current research. To give a few examples, on the one hand, openness and conscientiousness (Hibbing et al., 2019), as well as

simplistic thinking moderated with the time spent on the questionnaire were found to predict ERS (Naemi et al., 2009). On the other hand, cultural factors such as ethnicity are also able to predict ERS (Arce-Ferrer, 2006; Clarke, 2000). Therefore, it is reasonable to suspect that personality traits and cultural factors might interact with the effect of response format on the frequency of extreme responses. Hence, it is important to note, that failing to consider these factors as moderators could have led to falsely suggesting no effect (Plieger & Reuter, 2020). Therefore, collecting more extensive data about participants and including these personal aspects as moderators in the analysis is likely to enhance the understanding of the interaction between personal factors and response formats on ERS.

      Moreover, it is also important to recognize the heterogeneity of the VAS and Likert scales. There could be other variations of these response scales that might lead to different effects (Joshi et al., 2015; Ogon et al., 1996). For instance, changing the orientation of the VAS (Ogon et al., 1996), altering the length or granularity of the Likert scale (Hui & Triandis, 1989; Kieruj & Moors, 2013), including labels at intermediate points of the VAS (García-Pérez & Alcalá-Quintana, 2022), or adding emoticons (He et al., 2022). More complicated scale designs could lead to an increase in ERS occurrence due to the connection established between effort and use of response heuristics (Funke & Reips, 2012; Kieruj & Moors, 2013; Krosnick, 1991). Alternatively, adding intermediate labels to VAS or Likert scale could simplify finding corresponding responses to the feelings of the participants (García-Pérez & Alcalá-Quintana, 2022). Similarly, using emoticons could also simplify the response process on the VAS and make accurate responding less cognitively demanding, making ERS less likely (He et al., 2022; Krosnick, 1991). Therefore, investigating the heterogeneity of VAS and Likert scales is crucial, as various modifications could yield different outcomes.

      In order to conduct a comparison of several variations, further research might also consider varying the design of the investigation. The currently used between-subject research design can provide very solid results, however, using a within-person repeated measurement research design might be the more efficient way of investigating the large number of variations of the response scales. Using multiple variations in one study would allow for a simultaneous comparison of many scales at once, thereby, also demanding fewer participants than the current research's between-subject designs does. However, caution must be exercised, as within-subject design can be prone to lower precision and greater susceptibility to biases, such as carry-over - or order effects (Charness et al., 2012; Nicholas et al., 2012).

Nonetheless, such a comprehensive exploration of various scale designs yields great potential for providing better guidelines and recommendations to researchers in designing their studies.

Strength and Limitations

The major strengths of this study are that it was the first ESM study to compare the effect of VAS and Likert scale on ERS and that it did so using the very robust between-subject design in a randomised controlled trial. While previous research has examined the psychometric properties of VAS and Likert scales separately, none has directly compared these two response formats in the context of ESM. Therefore, this study fills an important research gap and offers valuable insights into the potential differences or similarities between VAS and Likert scales in triggering ERS. Furthermore, within the broader scope of research comparing VAS and Likert scales, also recognising studies outside of the context of ESM, this and related studies from this research project contribute as the first comparison focusing on ERS.

Nonetheless, certain limitations in this study may have impacted the validity of the results. Firstly, the convenience sampling method used resulted in a non-representative sample. The majority of participants were female, German, with an average age of 29.71. This limited diversity in the sample restricts the generalizability of the findings, especially in light of the importance of cultural and personal influences on ERS (Arce-Ferrer, 2006; Clarke, 2000; Hibbing et al., 2019; Naemi et al., 2009). Additionally, the inclusion of predominantly university students, likely from the University of Twente, introduces a potential familiarity bias towards the Likert scale, presenting the potential that the VAS scale might have been more difficult for participants to complete, which can be influential for the occurrence of ERS and might have also impacted retention rates (Hibbing et al., 2019; Krosnick, 1991).

Moreover, compared to other ESM studies, a severely larger number of participants had to be excluded due to a high number of non-responses (Vachon et al., 2019). While it was necessary to remove participants with insufficient response frequency to ensure data quality, it may have inadvertently resulted in the underrepresentation of individuals with specific personality characteristics, such as low motivation and conscientiousness, which are associated with a higher occurrence of ERS (Hibbing et al., 2019; Krosnick, 1991). Additionally, these high exclusion rates also contributed to overall small sample size, resulting in an underpowered analysis.

Implications

      While the findings of the study must be seen with its limitations in mind, the results bring along multiple implications for the design of ESM studies and the research into response formats. For the practical use of response scales, the results of this study cautiously point towards suggesting an interchangeability of Likert scale and VAS. This would give researchers the freedom to base their research design decisions on the specific characteristics of their sample without worrying about an effect on ERS. Regarding theoretical research, these results tentatively suggest that findings of studies that used Likert scale and VAS are comparable without a need to consider any adverse influences of ERS. This also implies that future research efforts could shift their focus towards exploring other aspects of response formats, such as different scale designs or biases that may influence participants' responses.

Conclusion

      This study represents the first investigation specifically designed to compare the impact of VAS and Likert scale on ERS within the context of ESM research. Findings showed no significant effect, indicating that the choice between VAS and Likert scales does not have an impact on ERS in ESM research. Therefore, it was suggested that ERS is likely a matter of personal and cultural factors. Nonetheless, certain limitations, such as the non-representative sample, potential exclusion bias and an uncommon method of measuring ERS, warrant caution in generalizing the findings.

References

Arce-Ferrer, A. J. (2006). An Investigation Into the Factors Influencing Extreme-Response Style. *Educational and Psychological Measurement*, *66*(3), 374–392. https://doi.org/10.1177/0013164405278575

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

Beal, D. J. (2015). ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Bishop, P. A., & Herron, R. L. (2015). Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International Journal of Exercise Science*, *8*(3), 297–302.

Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The 'I' in extreme responding. *Journal of Consumer Psychology*, *26*(4), 510–523. https://doi.org/10.1016/j.jcps.2016.03.002

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*(1), 147–168. https://doi.org/10.1016/0749-5978(91)90010-Q

Clarke, I. III. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality.*, *15*(1), 137–152.

Couclelis, H., Golledge, R. G., Gale, N., & Tobler, W. (1987). Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology*, *7*(2), 99–122. https://doi.org/10.1016/S0272-4944(87)80020-8

Csikszentmihalyi, M. (2014). *Flow and the Foundations of Positive Psychology*. Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8

Dejonckheere, E., & Erbas, Y. (2022). Designing an Experience Sampling study. In *The Open Handbook for Experience Sampling Methodology* (pp. 33–70).

Eisele, G., Kasanova, Z., & Houben, M. (2022). Questionnaire Design and Evaluation. In *The Open Handbook for Experience Sampling Methodology* (pp. 71–89).

Ethica Data Services Inc. (2023). *Ethica* (No. 651). https://ethicadata.com

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Funke, F., & Reips, U.-D. (2012). Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods*, *24*(3), 310–327. https://doi.org/10.1177/1525822X12444061

García-Pérez, M. A., & Alcalá-Quintana, R. (2022). Accuracy and precision of responses to visual analog scales: Inter- and intra-individual variability. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02021-0

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *The Public Opinion Quarterly*, 328–351.

Gries, K., Berry, P., Harrington, M., Crescioni, M., Patel, M., Rudell, K., Safikhani, S., Pease, S., & Vernon, M. (2018). Literature review to assemble the evidence for response scales used in patient-reported outcome measures. *Journal of Patient-Reported Outcomes*, *2*(1), 41. https://doi.org/10.1186/s41687-018-0056-3

Guyatt, G. H., Townsend, M., Berman, L. B., & Keller, J. L. (1987). A comparison of Likert and visual analogue scales for measuring change in function. *Journal of Chronic Diseases*, *40*(12), 1129–1133. https://doi.org/10.1016/0021-9681(87)90080-4

He, S., Renne, A., Argandykov, D., Convissar, D., & Lee, J. (2022). Comparison of an Emoji-Based Visual Analog Scale With a Numeric Rating Scale for Pain Assessment. *JAMA*, *328*(2), 208. https://doi.org/10.1001/jama.2022.7489

Hibbing, M. V, Cawvey, M., Deol, R., Bloeser, A. J., & Mondak, J. J. (2019). The Relationship Between Personality and Response Patterns on Public Opinion Surveys: The Big Five, Extreme Response Style, and Acquiescence Response Style. *International Journal of Public Opinion Research*, *31*(1), 161–177. https://doi.org/10.1093/ijpor/edx005

Hui, C. H., & Triandis, H. C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Janssens, M., Eshuis, J., Peeters, S., Lataster, J., Reijnders, J., Enders-Slegers, M.-J., & Jacobs, N. (2020). The Pet-Effect in Daily Life: An Experience Sampling Study on Emotional Wellbeing in Pet Owners. *Anthrozoös*, *33*(4), 579–588. https://doi.org/10.1080/08927936.2020.1771061

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Kieruj, N. D., & Moors, G. (2013). Response style behavior: question format dependent or personal style? *Quality & Quantity*, *47*(1), 193–211. https://doi.org/10.1007/s11135-011-9511-4

Klein, S. B. (2015). What memory is. *WIREs Cognitive Science*, *6*(1), 1–38. https://doi.org/10.1002/wcs.1333

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kuhlmann, T., Dantlgraber, M., & Reips, U.-D. (2017). Investigating measurement equivalence of visual analogue scales and Likert-type scales in Internet-based personality questionnaires. *Behavior Research Methods*, *49*(6), 2173–2181. https://doi.org/10.3758/s13428-016-0850-x

Laerhoven, H. van, Zaag-Loonen, H. van der, & BHF, D. (2004). A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatrica*, *93*(6), 830–835. https://doi.org/10.1080/08035250410026572

Lukacz, E. S., Lawrence, J. M., Burchette, R. J., Luber, K. M., Nager, C. W., & Galen Buckwalter, J. (2004). The use of Visual Analog Scale in urogynecologic research: A psychometric evaluation. *American Journal of Obstetrics and Gynecology*, *191*(1), 165–170. https://doi.org/10.1016/j.ajog.2004.04.047

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Myin-Germeys, & Kuppens. (2022). Sampling Methods, an Introduction. In Myin-Germeys, Kuppens,. In *The Open Handbook for Experience Sampling Methodology* (pp. 7–18).

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality Predictors of Extreme Response Style. *Journal of Personality*, *77*(1), 261–286. https://doi.org/10.1111/j.1467-6494.2008.00545.x

Nicholas, J. M., Grieve, A. P., & Gulliford, M. C. (2012). Within-person study designs had lower precision and greater susceptibility to bias because of trends in exposure than cohort and nested case–control designs. *Journal of Clinical Epidemiology*, *65*(4), 384–393. https://doi.org/10.1016/j.jclinepi.2011.09.004

Ogon, M., Krismer, M., Söllner, W., Kantner-Rumplmair, W., & Lampe, A. (1996). Chronic low back pain measurement with visual analogue scales in different settings. *Pain*, *64*(3), 425–428. https://doi.org/10.1016/0304-3959(95)00208-1

Plieger, T., & Reuter, M. (2020). Stress & executive functioning: A review considering moderating factors. *Neurobiology of Learning and Memory*, *173*, 107254. https://doi.org/10.1016/j.nlm.2020.107254

R Core Team. (2018). *R: A Language and Environment for Statistical Computing (4.3.0)*. R Foundation for Statistical Computing. https://www.R-project.org/

Sealed Envelope Ltd. (n.d.). *Randomisation and Online Databases for Clinical Trials*. Retrieved 24 June 2023, from sealedenvelope.com

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10.1145/3123988

van der Linden, K., Simons, C., Viechtbauer, W., Ottenheijm, E., van Amelsvoort, T., & Marcelis, M. (2021). A momentary assessment study on emotional and biological stress in adult males and females with autism spectrum disorder. *Scientific Reports*, *11*(1), 14160. https://doi.org/10.1038/s41598-021-93159-y

Versluis, A., Verkuil, B., Lane, R. D., Hagemann, D., Thayer, J. F., & Brosschot, J. F. (2021). Ecological momentary assessment of emotional awareness: Preliminary evaluation of psychometric properties. *Current Psychology*, *40*(3), 1402–1410. https://doi.org/10.1007/s12144-018-0074-6

Viechtbauer, W. (2022). Structuring, Checking and Preparing the Data. In l. Myin-Germeys & P. Kuppens (Eds.), *The Open Handbook for Experience Sampling Methodology* (pp. 137–150).

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wickham, & Hadley. (2016). *Ggplot2: Elegant graphics for data analysis* (No. 2). Springer International Publishing.

Appendix A

Enrolment Information Sheet: Likert Condition

Dear participant,

Thank you for your participation in the study on mental health in daily life. We are contacting you because you kindly agreed to participate in this study for the bachelor psychology at the University of Twente.

Brief summary of the project

The study you are participating in is a daily diary study. With this study we want to investigate how people feel and react to events in their day-to-day lives. By asking a few questions at several moments throughout the day, we get an insight in behavior of people in their everyday environment, which is necessary if we want to understand how people behave and feel in daily life. You will receive a notification on 10 random moments a day to answer a short questionnaire which will take about 1 minute to complete. We ask you to do this for 7 days in a row. The first questionnaire will be send on Monday morning, the 17th of April. Of course, there are situations in which it is not possible to fill it out (such as when you are driving), but to get a good overview of your daily life it is important that you fill out as many of these questionnaires as possible. In addition to these short questionnaires, you will receive one questionnaire in the beginning of the study that takes about 20 minutes to complete. It's important that you complete this questionnaire as well.

How to get ready to participate

Before continuing, make sure to download the Ethica application on your smartphone. Clicking on the following links on your smartphone will bring you the app store.

Android:

https://play.google.com/store/apps/details?id=com.ethica.logger&hl=en_US&gl=US&pli=1

IOS: https://apps.apple.com/nl/app/ethica/id1137173052

Then follow the these steps:

- Open the Ethica application on your phone. **Please make sure to allow push notifications for the Ethica app on your phone!**
- Click on "Sign up" and create an account.

- After you signed up in Ethica, login in to the Ethica application using your username and password.
- After logging in, click on the following link on your phone:

https://ethicadata.com/study/2349/

- Alternatively, you can also directly enter the registration code **2349** in the Ethica application.
- On the next window click on "Register" to enroll in the study.
- The study should now be set up and you will receive the first questionnaire next Monday.

Contact details

This study is part of a larger project with many researchers involved. If you have any questions, you can contact one of the following students who are involved in data collection or the supervisors. The contact details can be found below.

Students

Simon Brune

Nick Delventhal

Jan Derksen

Gina Haccou

Samuel Pietsch

Aleksandra Popovic

Lea Staudigel

Nina Zarrin Tigh

Supervisors

Jannis Kraiss

Thomas Vaessen

Thank you for participating in this study. Your contribution is greatly appreciated.

Kind regards, also on behalf of the whole study team.

Appendix B

Enrolment Information Sheet: VAS Condition

Dear participant,

Thank you for your participation in the study on mental health in daily life. We are contacting you because you kindly agreed to participate in this study for the bachelor psychology at the University of Twente.

Brief summary of the project

The study you are participating in is a daily diary study. With this study we want to investigate how people feel and react to events in their day-to-day lives. By asking a few questions at several moments throughout the day, we get an insight in behavior of people in their everyday environment, which is necessary if we want to understand how people behave and feel in daily life. You will receive a notification on 10 random moments a day to answer a short questionnaire which will take about 1 minute to complete. We ask you to do this for 7 days in a row. The first questionnaire will be send on Monday morning, the 17th of April. Of course, there are situations in which it is not possible to fill it out (such as when you are driving), but to get a good overview of your daily life it is important that you fill out as many of these questionnaires as possible. In addition to these short questionnaires, you will receive one questionnaire in the beginning of the study that takes about 20 minutes to complete. It's important that you complete this questionnaire as well.

How to get ready to participate

Before continuing, make sure to download the Ethica application on your smartphone. Clicking on the following links on your smartphone will bring you the app store.

Android:

https://play.google.com/store/apps/details?id=com.ethica.logger&hl=en_US&gl=US&pli=1

IOS: https://apps.apple.com/nl/app/ethica/id1137173052

Then follow the these steps:

- Open the Ethica application on your phone. **Please make sure to allow push notifications for the Ethica app on your phone!**
- Click on "Sign up" and create an account.

- After you signed up in Ethica, login in to the Ethica application using your username and password.
- After logging in, click on the following link on your phone:

https://ethicadata.com/study/1296/

- Alternatively, you can also directly enter the registration code **1296** in the Ethica application.
- On the next window click on "Register" to enroll in the study.
- The study should now be set up and you will receive the first questionnaire next Monday.

Contact details

This study is part of a larger project with many researchers involved. If you have any questions, you can contact one of the following students who are involved in data collection or the supervisors. The contact details can be found below.

Students

Simon Brune

Nick Delventhal

Jan Derksen

Gina Haccou

Samuel Pietsch

Aleksandra Popovic

Lea Staudigel

Nina Zarrin Tigh

Supervisors

Jannis Kraiss

Thomas Vaessen

Thank you for participating in this study. Your contribution is greatly appreciated.

Kind regards, also on behalf of the whole study team.

Appendix C

Informed Consent

Dear participant,

Thank you for your participation in this study.

Brief summary of project

The study is using the Experience Sampling Method (ESM) to obtain data. This means that 10 times a day there will be a prompt to answer a questionnaire containing about 20 items, which will take about 1 minute to complete. The questions regard your psychological well-being in the specific moment you are receiving the questionnaire and the time in-between questionnaires. It is important to fill out as many questionnaires as possible to ensure the success of the project.

To participate in this study, we need to ensure that you understand the nature of the research, as outlined in the participant information sheet. Please confirm at the bottom of the page to indicate that you understand and agree to the following conditions:

- I confirm that I have read the participant information sheet for this study. I have had the opportunity to consider the information, ask questions, and have had these answered satisfactorily
- I understand that to take part in this study, I should
- Be at least 18 years old
- Possess a basic level of English
- I understand that personal data about me will be collected for the purposes of the research study including age, gender, nationality, level of education, current studies, and primary occupation, and this data will be processed completely anonymous and in accordance with data protection regulations.
- I understand that taking part in this study involves that I will be filling in 10 questionnaires every day for one week.
- I am voluntarily taking part in this research, and I know that I can stop the research at any time without giving any reason, without my rights being affected
- I don't expect to receive any benefit or payment for my participation.
- I understand that I am free to contact the researchers or supervisor with any questions I may have in the future.

- I understand that the data collected in this study will be anonymized, and only be used for academic purposes i.e., writing a thesis for the bachelor and/or master.
- I understand that personal data that will be collected within this study will not be shared with anyone other than the study team.
- I agree to take part in this study.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee/domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-hss@utwente.nl

Study contact details for further information:

Students

Simon Brune

Nick Delventhal

Jan Derksen

Gina Haccou

Samuel Pietsch

Aleksandra Popovic

Lea Staudigel

Nina Zarrin Tigh

Supervisors

Jannis Kraiss

Thomas Vaessen