Endpoint-Only-Labelling versus Full-Labelling - The Effect of Scale Design on ESM Measures

Lorena Haase

Faculty of Behavioural, Management, and Social Sciences

Department of Psychology, University of Twente

Bachelor Thesis

First Supervisor: Dr. Thomas R. Vaessen

Second Supervisor: Mieke J. M. van Bergen

January 27 2023

Abstract

Background: ESM is a research method growing in popularity as it can reliably assess various momentary processes such as emotion dynamics which can forecast mental well-being. Studies with retrospective questionnaires have found that when endpoint-only-labelling and full-labelling are compared, they can show different study outcomes. Yet the influence of scale design on results of ESM studies has not been investigated although it is important to correctly measure the concept of emotional variability, which is measured by assessing negative affect (NA) and positive affect (PA). If labelling influences scores, this would lead to different conclusions about psychosocial processes. Thus, the current study aimed at examining the effect of full-labelling compared to endpoint-only-labelling on NA and PA measures of mean, distribution, and emotional variability.

Method: Participants were gathered by convenience sampling. As a study method ESM was used, where participants were prompted to fill out ten questionnaires per day for seven consecutive days in the app Ethica. The questionnaires measured NA and PA. Linear mixed model analyses with condition as a fixed effect, scores of either NA or PA as dependent variable, and participant as random effect were used to examine differences between mean measures of PA and NA. Differences between distributions were investigated by comparing the confidence intervals of kurtosis and skewness measurements. Finally, regression analyses with the within-person standard deviation of PA or NA as the dependent variable and condition as a predictor were carried out to examine group differences of emotional variability.

Results: There was no significant difference between conditions for the mean of NA and PA, or emotional variability. Regarding distribution measures, the significant differences found between conditions were the skewness of PA (full-labelling = -0.25, 95% CI [-0.40, -0.09], endpoint-only-labelling = 0.08, 95% CI [-0.08, 0.16]) and kurtosis of NA (full-labelling = 2.24, 95% CI [1.93, 2.55], endpoint-only-labelling = 1.06, 95% CI [0.75, 1.37]).

Conclusion: This study was the first to investigate the effect of scale design on ESM measures but as most differences between conditions were insignificant, endpoint-only-labelling seems to be a valid way of designing scales used in ESM studies. Nevertheless, the effect of scale design on ESM measures needs to be investigated more in future studies and whether the effect of habituation might be significant in that regard.

Introduction

One of the evolving research methodologies in psychology is the Experience Sampling Method (ESM) as it has been growing in popularity for the past few decades (Xu & Wang, 2019). ESM is a structured self-report diary technique that can assess mood, certain symptoms such as pain or fatigue, and the context of appraisals in daily life (van Os et al., 2017). Participants usually need to fill out a short questionnaire multiple times a day for several days. Through this technique, researchers can study subjective experiences as they occur in everyday environments while maintaining ecological validity (Brunswick as cited by Csikszentmihalyi & Larson, 2014; Myin-Germeys et al., 2018). Thus, ESM is an attempt to be as objective as possible about subjective experiences like emotions and still capture context-related information to identify and analyse possible patterns (Csikszentmihalyi & Larson, 2014). Although a vast amount of ESM research has been conducted (e.g., ScienceDirect shows 6,640 results when searching for ("experience sampling" OR "momentary assessment")), there has been no extensive research on the influence of different scale designs on ESM research results as has been done with cross-sectional designs. As one of the most relevant topics in ESM research is the investigation of emotion dynamics by measuring positive and negative affect, the influence of labelling on the measure of those emotion dynamics should be considered. Positive affect (PA) and negative affect (NA) are composite scores. They are comprised of averaging several items (van Os et al., 2017) and will be the relevant measures in this report.

ESM Scale Design

Previous studies using retrospective questionnaires indicate that the labelling of anchor points affects the interpretation of such by both researchers and participants (Johnson et al., 2005; Weng, 2004). Labelling also influences response styles, which has an effect on the resulting data (Krosnick, 1991; Swain et al., 2008; Weijters et al., 2010). As scale design could influence conclusions about psychosocial processes, the necessity emerges to investigate whether labelling would have a similar influence on ESM data. In this thesis, the focus will be on the difference between labelling only the endpoints of a scale with verbal labels (endpoint-only-labelling) and labelling all anchor points with verbal labels (full-labelling). Endpoint-only-labelling is mostly used in ESM research as scales are mostly presented on a small screen, such as from a smartphone, making screen size a potential issue when designing and selecting scales. This limitation can hinder researchers from labelling each anchor point on a Likert scale to make the scale more readable for the respondents (Väätäjä & Roto, 2010). Although there is a practical obstacle to making full-labelling in ESM questionnaires more difficult, the effect of different labelling should be investigated as full-labelled scales are generally recommended even in settings where small displays are the means of presentation (Gummer & Kunz, 2021).

Effect of Labelling

Compared to ESM, with traditional data-gathering techniques such as retrospective questionnaires, full-labelling is generally preferred by researchers and respondents (Krosnick & Fabrigar, 1997). Studies have shown that even subtle changes in the layout and appearance of rating scales can affect responses (Matejka et al., 2016; Reips, 2010; Schwarz et al., 1991; Smyth et al., 2006) and response times (Tourangeau et al., 2007), which consequently influences study results. This influence could be due to full-labelling generally providing more information on how to interpret scales to item respondents (Johnson et al., 2005; Weng, 2004), which can be important for identifying scale anchors. In a study by Arce-Ferrer (2006) which investigated the effect of labelling on response styles, only one-fifth of respondents could label all anchor points correctly. Furthermore, labelling can be influential regarding the reliability of a questionnaire. While some studies suggest that endpoint-only-labelling might provide higher reliability when compared to full-labelling (Andrews & Crandall, 1976; Rodgers et al., 1992), most studies show that reliability increases with full-labelled scales when compared to endpoint-only-labelling (Krosnick & Berent as cited by Schwarz et al., 1991; Menold et al., 2014; Saris & Gallhofer as cited by DeCastellarnau, 2018). Additionally, labelling choices can affect distributions of measurements and response distributions may differ depending on the chosen labels (French-Lazovik & Gibson, 1984; Weijters et al., 2010). Despite this evidence, research on the effect of different design choices, such as full-labelling compared to endpoint-only-labelling on data quality and quantity in ESM questionnaires, has not been investigated yet.

Influence on Response Styles

An important factor regarding the influence of labelling is response bias. According to Greenleaf (1992), response category labels may influence the level of response bias. Response bias is a response style that occurs when there is a systematic tendency to respond to questionnaire items based on something different than what the items are designed to measure (Paulhus, 1991). Tourangeau et al. (2007) found that endpoint-only-labelling might lead to a generally increased use of response styles as a form of heuristics. Three different response styles exist: Net Acquiescence Response Style (NARS), Extreme Response Style (ERS), and Misresponse to

Reversed Items (MR). Only ERS and NARS will be relevant as in ESM affect assessment typically no reversed items are used.

ERS is the tendency to choose only or disproportionally often the extreme endpoints of the scale, which affects the spread of the observed data (Baumgartner & Steenkamp, 2001; Greenleaf, 1992; Hurley, 1998). Fully labelled scales, however, seem to reduce ERS (Eutsler & Lang, 2015; Moors et al., 2014) due to the increased salience and attractiveness of intermediate options, resulting in a lower standard deviation (Krosnick, 1991; Swain et al., 2008). Thus, fully labelled scales might also lead to a lower standard deviation and a higher kurtosis of the distribution in ESM questionnaires when compared to endpoint-only labelled scales.

NARS describes the tendency of respondents to show greater acquiescence (tendency to agree) rather than disacquiescence (tendency to disagree) with items, regardless of content (Baumgartner & Steenkamp, 2001; Greenleaf, 1992; Van Herk et al., 2004). According to a study by Weijters et al. (2010), this tendency to agree increases with fully labelled scales due to the added clarity of full-labelling, which strengthens the effect of positivity bias (Tourangeau et al. as cited by Weijters et al., 2010), the tendency to report positive views of reality (Hoorens, 2014). Overall, NARS could lead to generally higher mean scores in ESM questionnaires with full-labelling compared to endpoint-only-labelling. Furthermore, NARS could influence distribution measures and lead to more skewed data for PA measures due to the increased tendency to agree with full-labelled scales compared to endpoint-only-labelled scales. For NA measures, however, the data is expected to be less skewed as NA measures usually do not show a normal distribution but rather one that is positively skewed (Crawford & Henry, 2004). Thus, the data should be less positively skewed with full-labelling compared to endpoint-only-labelling.

Emotion Dynamics and the Importance of Labelling

Until now, endpoint-only-labelling was the standard in ESM questionnaires measuring emotion dynamics. Although previous ESM studies might still be relevant and valid as the results of these studies generally show high reliability and high validity (Csikszentmihalyi & Larson, 2014; Hektner et al., 2007), the effect of design choices should be investigated as they seem to affect responses of participants and therefore possibly the measurement of emotion dynamics. In emotion dynamics, three concepts are important: emotional variability, emotional stability, and emotional inertia. This paper will focus on emotional variability.

Emotional variability is “the range or amplitude of someone’s emotional states across time” (Houben et al., 2015, p. 902) and seems to be highly associated with psychological well-being.

Excessive fluctuations in emotion are maladaptive and can forecast decreased well-being (Holmes et al., 2016; Houben et al., 2015; van Zutphen et al., 2015) or even psychological illnesses (Thompson et al., 2017). Thus, measuring emotional variability accurately is essential. Labelling choices could influence measures of variability due to the aforementioned response styles. As NARS seems to increase with full-labelling compared to endpoint-only-labelling, variability could decrease as the tendency to select 'agree'-responses might be higher. Furthermore, the within-standard deviation of participants is used to compute scores of emotional variability. As measures of standard deviation seem to be lower by using full-labelled scales compared to endpoint-only-labelled scales (Krosnick, 1991; Swain et al., 2008), this further underpins the expectation that variability scores should be lower with full-labelling when compared with endpoint-only-labelling.

The Current Study

The following study set out to investigate the effect of full-labelling compared to endpoint-only-labelling in ESM questionnaires on descriptive and psychometric variables such as mean, skewness and kurtosis of the data as well as the measure of emotional variability. Six research hypotheses were formulated and will be tested in this paper:

H1: Mean NA will be higher in the full-labelling condition vs endpoint-only-labelling condition.

H2: The distribution of NA scores will be more skewed and have a lower kurtosis in the endpoint-only-labelling condition compared to the full-labelling condition.

H3: Mean PA will be higher in the full-labelling condition vs endpoint-only-labelling condition.

H4: The distribution of PA scores will be less skewed and have a lower kurtosis in the endpoint-only-labelling condition compared to the full-labelling condition.

H5: Emotional variability will be lower for NA in the full-labelling condition vs endpoint-only-labelling condition.

H6: Emotional variability will be lower for PA in the full-labelling condition vs endpoint-only-labelling condition.

Methods

Design

The study was conducted using ESM. To meet the objectives of this study, a between-subjects design was performed with two conditions (endpoint-only-labelling vs full-labelling).

This means the questionnaire participants had to fill out multiple times a day either had only the endpoints labelled or all possible options.

Participants

The research group was aiming for at least 60 participants in total, 30 for each condition, as the researchers oriented themselves among other ESM studies (e.g., van Berkel et al., 2017; van Berkel et al., 2019) and decided that at least 30 participants per condition were needed. To account for possible insufficient data, the researchers ensured that more participants were recruited. They were recruited in October 2022, one month before the experiment started. The participants were gathered through SONA, where psychology students from the University of Twente participate in studies in exchange for credits they need to collect during their bachelor studies. In total, students need to collect 15 credits. By taking part in this study, they could collect two credits. Furthermore, the researchers used convenience sampling and reached out to potential participants in their social environment. Those participants did not receive any compensation. All participants needed to fulfil the inclusion criteria of being 18 or older, owning a smartphone on which the application Ethica could be downloaded, and speaking a sufficient level of English, as the whole study was conducted in English. Before the actual study could start, ethical approval was obtained through the BMS Ethics Committee of the University of Twente. Furthermore, participants gave informed consent (see Appendix A) prior to participation.

Randomization

Before recruiting participants, the researchers ensured that the participants were assigned to one of the two conditions randomly. Stratified randomization was implemented to account for possible similarities of participant characteristics due to being sampled from the researcher's social environment. An initial randomization schedule was generated before recruiting participants for 20 participants per researcher and up to 60 participants from SONA. The conditions were divided into blocks of six participants each, where half of those participants were assigned to the first condition and the other half to the second one. Participants' SONA number or e-mail address was put into the randomization sheet as soon as they signed up for the study.

Procedure

The ESM study was conducted with the online tool Ethica which enables the design of online research studies (Ethica Data Services Inc, 2022). The app needed to be downloaded by the participants on their phones before the start of the study. Participants received an email with the necessary study information three days before the study started (see Appendix B). The email included a link to Ethica and the registration code that was needed to register for the study in the app. Furthermore, important information for the participants about ESM as a method was mentioned. For example, the information included the number of questionnaires, that as many questionnaires as possible need to be filled out, and that the participant's life should not be changed for answering the questionnaire (e.g., changing their schedule, using their phone while driving to answer a questionnaire). Everything was written in an easily understandable way without using any jargon.

All participants started the study at the same time on Monday 7-11-2022 at 7:30 and ended on Sunday 13-11-2022 at 22:30. First, participants were asked to fill out a baseline questionnaire in the app, which assessed well-being, as well as demographical data. Afterwards, the ESM period started where participants were prompted to fill out a questionnaire ten times a day for seven days, the duration of the study, which resulted in a total of 70 questionnaires. The notifications to fill out the ESM questionnaire were at random moments during one of the ten 90-minute blocks the day was divided into (see Table 1). If the questionnaire was not filled out within 15 minutes after the notification of the app, the questionnaire expired and could not be answered anymore.

Table 1

The Schedule of the Study for all Days, Including Relevant Variables, Points in Time, Expire time and Notification for the Different Questionnaires

Day	Questionnaire	Relevant variables	Points in time	Expire time	Notifications
1	Demographics Baseline well-being	All	1	No	1
1-7 (7 days)	Daily questionnaire	PA and NA	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	1 (10 in total/each day)

Measures

Demographics

All participants were asked for their age, gender, nationality, occupation, and level of education. Participants who would receive SONA points for participating in the study were asked to mention their SONA identification number.

ESM questionnaire – Positive and Negative Affect

To measure PA and NA, a set of eight items was used which was based on the PANAS scale (Watson et al., 1988). Four items measured PA and NA each. The items were: "How cheerful / enthusiastic / satisfied / relaxed do you feel right now?" for PA. For measuring NA, the items were: "How anxious / irritable / down / guilty / stressed do you feel right now?". The average of the items per scale was calculated to compute scores of PA and NA

All affect items were measured on a 7-point Likert scale for both conditions. For the endpoint-only-labelling condition, the scale was labelled only at the endpoints (i.e., 1 and 7) with 1 = *not at all* and 7 = *extremely* and otherwise consisted of numbers (i.e., 2, 3, 4, 5, and 6). The scale for the full-labelling condition was labelled at all points as follows: 1 = *not at all*, 2 = *very slightly*, 3 = *slightly*, 4 = *moderately*, 5 = *much*, 6 = *very much*, and 7 = *extremely*. The endpoint-only-labelling condition had a Cronbach's alpha of α = .885 for the scale of PA and a Cronbach's alpha of α = .818 for the scale of NA. The full-labelling condition had a Cronbach's alpha of α = .927 for the scale of PA and a Cronbach's alpha of α = .880 for the scale of NA.

Data Analysis

To analyse the data, it was imported to IBM SPSS Statistics Version 28.0.1.0 (IBM Corp, 2022). Afterwards, the data was checked, and participants were removed whose data showed potential errors, such as having too many data points due to an error in the app Ethica during data collection. Furthermore, participants who filled out less than one-third of the ESM questionnaires were removed. Descriptive statistics were examined to gain a general understanding of the participant's demographic data (see Table 2).

To analyse the influence of condition on the measures of PA and NA, a series of Linear Mixed Models analyses were conducted since those types of analyses can account for random effects of participants. Additionally, mixed models are an appropriate tool to analyse complex datasets containing repeated observations per person as mixed models can deal with the resulting interdependence of data points. Furthermore, mixed models are suitable to deal with high amounts of missing data (Myin Germeys & Kuppens, 2021). To specify whether a result would be significant, a p-value of 0.05 was used as a significance level for the analyses. In the mixed model that was run, the condition was a fixed effect and treated as a dummy variable (0 = endpoint-only-labelling condition, 1 = full-labelling condition), the means of either PA or NA were the dependent variable, and participant was included as a random effect. To ensure that possible effects could be reliably attributed to the factor condition, a sensitivity analysis was conducted, in which those variables were included in the mixed model as random effect as education, gender and age were shown to possibly influence response styles and therefore results (Meisenberg & Williams, 2008).

To test the hypotheses on the distributions of NA and PA, the descriptive statistics of the mean PA and NA measures were inspected to check the skewness and kurtosis. Based on whether there was an overlap of the confidence intervals, it was decided whether the difference between the conditions would be significant. If confidence intervals overlapped, the difference was decided to be not significant.

To test the influence of the condition on the measure of emotional variability, the within-person standard deviation was calculated first. Afterwards, a regression analysis was run with the condition as the independent variable and the person-centred standard deviation as the dependent variable.

Results

Demographics

In total, 88 participants took part in the study. Data from two participants had to be excluded due to data that was missing because of an error in the application, and data from 36 participants were excluded as less than 33% of ESM questionnaires were responded to, making the data possibly unrepresentative for the daily life of participants (Myin-Germeys & Kuppens, 2021, p. 146). A final three participants were removed due to too many data points caused by an error in the app Ethica during the data collection period, bringing the final sample to 47 participants (M_{Age} = 24.39, *SD* = 7.69, range = 18 to 59). The majority of participants were female (59.1%), Dutch (36.4%) or German (54.5%), studying (75.0%), and had a high school diploma as their highest education level (52.3%). On average, participants filled out 58.8% (*SD* = 49.23) of the questionnaires with a mean completion time of 75.36 seconds (*SD* = 76.20).

Explorative analysis

A few exploratory analyses were conducted to get a first understanding of the data and to potentially understand patterns in the data better. To compare the individual mean scores between groups, a t-test was conducted for PA. The effect was not significant $t(45)$, $p = .893$. As the data was non-normally distributed for scores of NA, a Mann-Whitney U test was computed with the individual mean scores. Measures of the full-labelling condition did not show significant differences compared to the endpoint-only-labelling condition $U = 248.00$, $Z = -0.596$, $p < .551$ (see Table 2). The difference between conditions for the mean completion time for the ESM questionnaire was not significant $t(1887)$, $p = .808$.

Hypothesis Testing

H1: Mean NA is higher in the full-labelling condition vs endpoint-only-labelling condition.

After running the linear mixed model, there was no significant difference between the groups $F(1,44.92) = 0.03$, $p = 0.892$. Adding the covariates age, gender, and occupation as random effects to the model did not significantly alter the results.

H2: The distribution of NA scores is more skewed and has a lower kurtosis in the endpoint-only-labelling group compared to the full-labelling group.

The skewness of NA was found to be 1.45, 95% CI [1.29, 1.61] in the full-labelling condition and 1.28, 95% CI [1.12, 1.43] in the endpoint-only-labelling condition, indicating that the distribution was right-skewed (see Figure 1). The confidence intervals of the distribution's skewness were overlapping, suggesting that the distributions' skewness in the endpoint-only-labelling condition does not significantly differ from the distribution's skewness in the full-labelling condition. The kurtosis of NA was found to be 2.24, 95% CI [1.93, 2.55] in the full-labelling condition and 1.06, 95% CI [0.75, 1.37] in the endpoint-only-labelling condition, indicating that the distribution was less heavy-tailed compared to the normal distribution. The confidence intervals of the two conditions did not overlap, suggesting that the kurtosis in the endpoint-only-labelling condition is significantly lower than in the full-labelling condition.

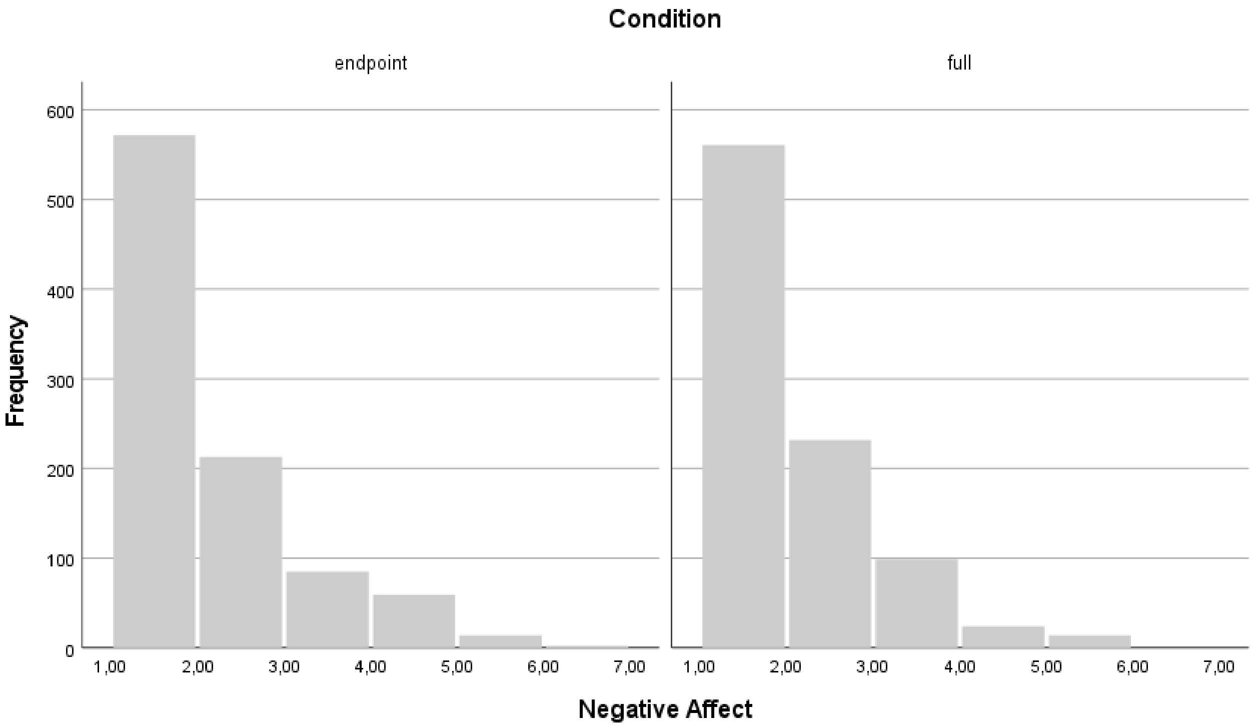
Table 2

Sample Descriptives

Variable	Endpoint-only-labelling (*n* = 23)	Full-labelling (*n* = 24)
Mean Age (*SD*)	24.55 (8.29)	24.05 (7.31)
	%	**%**
Gender		
Female	63.6	54.5
Male	31.8	40.9
Other	4.5	4.5
Nationality		
German	63.3	45.5
Dutch	27.3	45.5
Other	9.1	9.1
Occupation		
Working	18.2	18.2
Student	50.0	50.0
Studying and Working	31.8	22.7
Self-employed	-	4.5
Not working	-	4.5
Education		
Middle School	9.1	4.5
High School	63.6	45.5
Bachelor	18.2	40.9
Master	4.5	4.5
Other	4.5	4.5
ESM measures		
Mean PA (*SD*)	4.02 (1.01)	4.22 (0.71)
Median NA	1.7	1.77
Mean Kurtosis NA	1.06, 95% CI [0.75, 1.37]	2.24, 95% CI [1.93, 2.55]
Mean Skewness NA	1.28, 95% CI [1.12, 1.43]	1.45, 95% CI [1.29, 1.61]
Mean Kurtosis PA	0.08, 95% CI [-0.08, 0.16]	-0.25 [-0.40, -0.09]
Mean Skewness PA	-0.47, 95% CI [-0.78, -0.16]	-0.25, 95% CI [-0.57, 0.06]
Mean *SD* NA	0.68	0.72
Mean *SD* PA	1.05	1.03

Figure 1

Histogram of NA Measures



H3: Mean PA will be higher in the full-labelling group vs endpoint-only-labelling group.

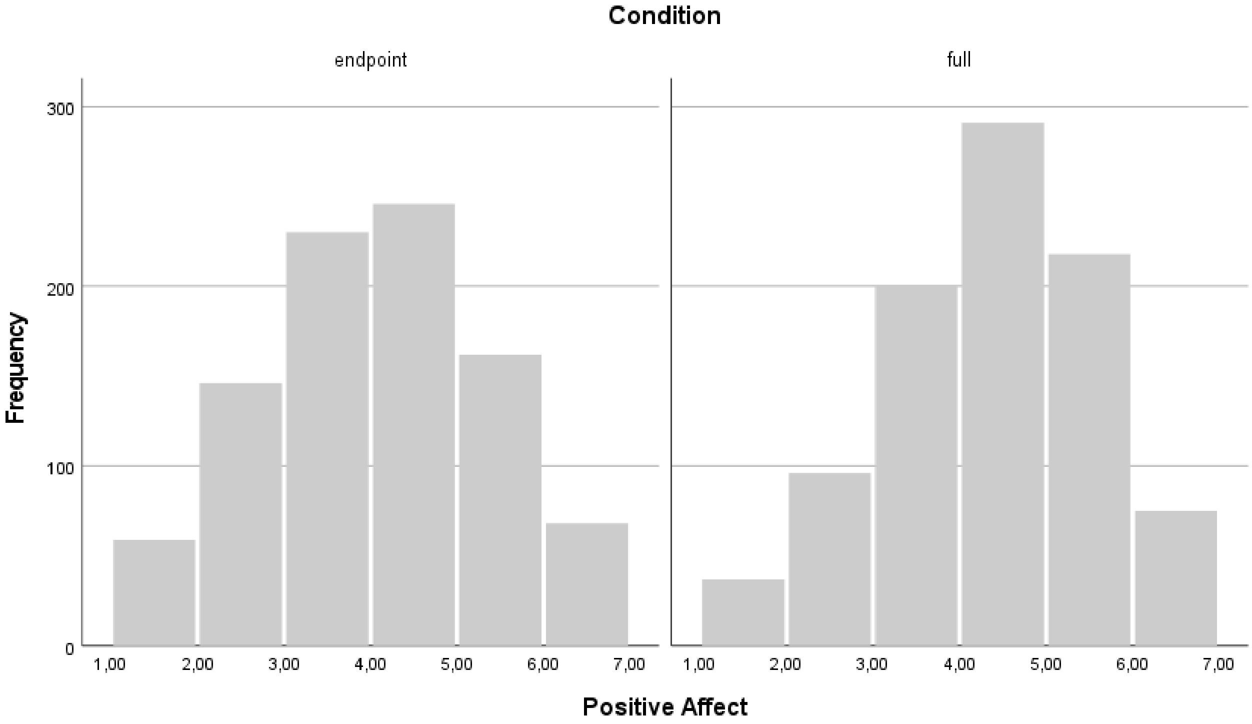
To test the third hypothesis, a linear mixed model was run with the condition as a fixed effect, participant as a random effect and PA scores as the dependent variable. The difference between the conditions was not significant $F(1,44.21) = -0.18$, $p = 0.492$. To account for possible influences of occupation, age and gender, a linear mixed model was run, including those variables as random effects. Adding these covariates to the model did not significantly alter the results.

H4: The distribution of PA scores will be less skewed and have a lower kurtosis in the endpoint-only-labelling condition compared to the full-labelling condition.

The skewness of PA was found to be -0.25, 95% CI [-0.40, -0.09] in the full-labelling condition and 0.08, 95% CI [-0.08, 0.16] in the endpoint-only-labelling condition, indicating that the distribution in the full-labelling condition was left-skewed and that the distribution was right-skewed in the endpoint-only-labelling condition (see Figure 2). The confidence intervals of the distribution's skewness were not overlapping, suggesting a significant difference in the skewness of the distribution between conditions. The kurtosis of PA was found to be -.25, 95% CI [-0.57, 0.06] in the full-labelling condition and -.47, 95% CI [-0.78, -0.16] in the endpoint-only-labelling condition, indicating that the distribution was less heavy-tailed compared to the normal distribution. The confidence intervals were overlapping; thus, no significant difference was found.

Figure 2

Histogram of PA Measures.



H5: Emotional variability will be lower for NA in the full-labelling condition vs endpoint-only-labelling condition.

Condition did not significantly predict the measure of emotional variability, $R^2 = .005$, $F(1, 45) = 0.23$, $p = .636$. The results are displayed in Table 3 and the regression line in Figure 3.

Table 3

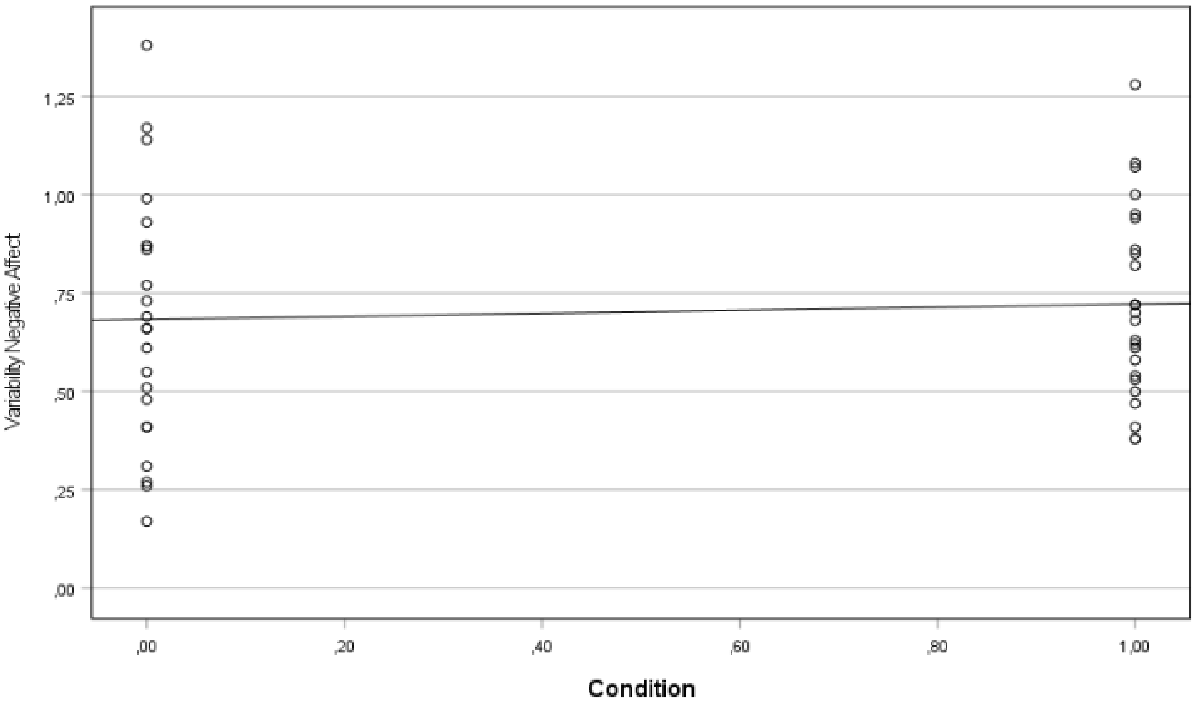
Coefficients Regression Analysis NA

Variable	B	95% CI	β	t	p
(Constant)	0.68	[0.57, 0.80]		11.67	<.001
Condition (endpoint = 0, full = 1)	0.04	[-0.12, 0.20]	0.07	0.48	.636

Note. N = 47, R^2adjusted = .005. CI = Confidence Interval for B.

Figure 3

Regression Line NA



H6: Emotional variability will be lower for PA in the full-labelling condition vs endpoint-only-labelling condition.

Condition did not significantly predict the measure of emotional variability, R^2 = .001, *F*(1, 45) = 0.06, *p* = .813. The results are displayed in Table 4 and the regression line in Figure 4.

Table 4

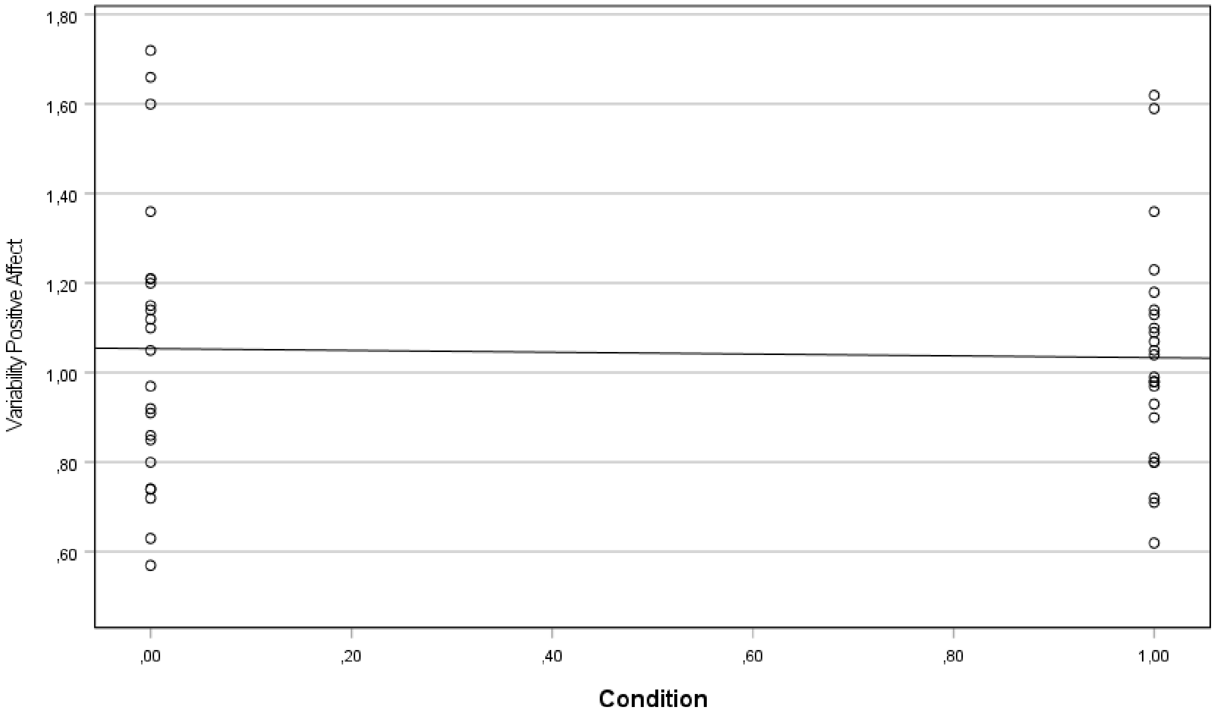
Coefficients Regression Analysis PA

Variable	B	95% CI	β	t	p
(Constant)	1.05	[0.93, 1.17]		17.79	<.001
Condition (endpoint = 0, full = 1)	-0.02	[-0.18, 0.14]	-0.04	-0.24	.813

Note. N = 47, R^{2}adjusted = .002. CI = Confidence Interval for B.

Figure 4

Regression Line PA



Discussion

This thesis examined the effect of scale design, specifically the influence of full-labelled scales compared to endpoint-only-labelled scales on measures of mean scores and distribution in ESM questionnaires as well as measures of emotional variability. Overall, there was no significant effect of scale design on means and variability of NA and PA measures. Furthermore, no significant differences were found between conditions for the skewness of NA and the kurtosis of PA. The only significant differences between conditions could be found for the kurtosis of NA and skewness of PA.

Interpretations

The means of NA did not differ significantly between conditions although full-labelled scales lead to supposedly higher means due to NARS as Weijters et al. (2010) found in their study where they compared endpoint-only-labelling with full-labelling. The different findings could be due to the difference in scale design as Weijters et al. (2010) used scales that were coded bipolar while the current study used a unipolar scale. Thus, unipolar scales might be more robust to effects of labelling regarding NARS.

Regarding distribution measures of NA, skewness was not significantly different between conditions despite the expectation that the distribution would be less skewed in the full-labelling condition due to positivity bias leading to more acquiescence (Tourangeau et al. as cited by Weijters et al., 2010) and possible effects of NARS (Weijters et al., 2010). Again, the scale design could have played a significant role as it might have for the mean of NA. Kurtosis in the endpoint-only-labelling condition was significantly lower than in the full-labelling condition, which fits with the second part of the hypothesis. The difference in kurtosis could be attributed to added clarity from the labels as the lowest option "1: not at all" was chosen less often in the full-labelling condition compared to the endpoint-only-labelling condition. This fits to the claim that ERS is reduced with full-labelled scales and therefore choosing extreme options less (Eutsler & Lang, 2015; Moors et al., 2014). Thus, the interpretation of the anchor points might have differed, which matches with a study by Arce-Ferrer (2006), where only one-fifth of respondents could interpret all anchor labels according to the intended meaning when no verbal labels were used.

The mean of PA did not differ between conditions. As for the mean of NA, this could be accounted to the different scale design that was used. Results showed a small but significant difference between conditions regarding the skewness of PA. While skewness was positive in the endpoint-only-labelling condition, it differed significantly in the full-labelling condition, where a negative and more skewed distribution was found. This finding fits the hypothesis that more skewness would be present in the endpoint-only-labelling condition. This result also fits the literature suggesting that full-labelling could lead to more acquiescence (Weijters et al., 2010) due to positivity bias (Tourangeau et al., as cited by Weijters et al., 2010). This poses the question of whether labels added clarity as the interpretation of the anchor points might have differed between groups, as it did in the previously mentioned study by Arce-Ferrer (2006). There was no significantly lower kurtosis in the endpoint-only-labelling condition that was anticipated based on results that showed a reduction of ERS with full-labelling compared to endpoint-only-labelling (Eutsler & Lang, 2015; Moors et al., 2014). Thus, labelling might be less important when choosing the endpoints for PA items and labelling more than endpoints might be more relevant for the intermediate options of PA measures as this led to a significantly different skewness.

No significant differences could be found between conditions for the measure of emotional variability (operationalized as the within-person standard deviation), which does not fit with results that showed that full-labelling leads to a lower standard deviation in retrospective questionnaires (Krosnick, 1991; Swain et al., 2008). The deviation from the expected differences between conditions for variability could be due to ESM being a method where questionnaires are filled out quickly, as there are several questionnaires per day. This might leave less time for participants to interpret the scales as the beeps generally interrupt participants' day-to-day-life. Thus, participants might be paying less attention to labels and more to the position of the anchor points. Furthermore, participants fill out the same questions every time and therefore get used to them. This is supported by the mean completion time as no significant difference between the two conditions could be found. This does not match previous literature on retrospective questionnaires where participants needed more time to fill out questionnaires with full-labelled scales compared to ones with endpoint-only labelled scales (Tourangeau et al., 2007) suggesting that habituation to items might be a factor that influences study results.

Overall, the differences between the significant conditions' distributional measures were relatively small, which makes it difficult to make reliable claims about significant differences, especially as no significant differences for other measures were found. Nevertheless, the results suggest a difference of the effect of labelling between PA measures and NA measures as the effect on distributional measures differed.

Implications

This study provides new insight into the effect of labelling on ESM measures. The findings suggest that previous ESM measures might have not been affected by design choices. Both scale designs, endpoint-only-labelling and full-labelling, seem to be valuable with regards to ESM measures. The only significant differences concerned the skewness of NA measures and the kurtosis of PA measures. Thus, one could consider whether using different labelling options for those measures could be valuable. As PA and NA are usually assessed together, using different scale designs for the respective items might lead to confusion. Based on the study results and taken together with the complicatedness of presenting full-labelled scales on small screens (Väätäjä & Roto, 2010), endpoint-only-labelling seems to be a suitable scale design for ESM questionnaires.

Strengths and Limitations

The study had some limitations that might constrain the generalizability of the study results. Since convenience sampling is the most accessible method to gather participants, most of the sample consisted of students. Therefore, the generalizability of the results is limited by the sample that was used in the study, as the participant characteristics may not be representative of the general population. Although occupation, age or education seemingly did not influence the results, it might be that there might not be enough participants with other characteristics to examine this effect accurately, as only 25% of participants were no students. Thus, to eliminate the effect of those demographic variables, the study was supposedly underpowered. Students might be more used to questionnaires and, thus, less prone to be influenced by scale design.

Nevertheless, there were some important strengths in conducting this research. The research was the first to investigate the effect of labelling choices on psychometric properties in ESM. Thus, this study is valuable regarding future choices concerning scale design in ESM research. Efforts were made to make reliable claims possible. As most of the sample was from the social environment of the researchers, the sample was randomized to prevent effects from being attributed to characteristics such as personality traits that are associated with the social environment of one of the researchers. Furthermore, regarding the number of participants, the researchers made sure to have enough participants for making reliable claims as they oriented themselves among existing ESM research (e.g., van Berkel et al., 2017; van Berkel et al., 2019).

Future Directions

For future research, the sample could be adjusted making the sample consist of participants with more diverse backgrounds. This way, the influence of for example education could be accounted for if it has an effect, and possible differences between education levels regarding scale designs could be investigated. If differences arise, those could be influential for study outcomes of past ESM studies. Next to sample characteristics, habituation was assumed to be a factor that could have possibly influenced the study results. Thus, the role of habituation in ESM concerning scale designs could be investigated by comparing whether the amount of times respondents fill out a questionnaire affects the results, and whether respondents would detect a change a change in scale design (e.g., by comparing endpoint-only-labelling vs full-labelling vs using both options alternating). Furthermore, the effect of labelling when using bipolar scales in ESM questionnaires could be investigated as these yielded significant differences in Weijter et al.'s (2010) study. Additionally, a comparison between unipolar and bipolar scales would be a field worth exploring as Schwarz et al. (1991) found that those can lead to significantly different results. If they show significant differences in ESM study outcomes, this will also influence conclusions about psychosocial processes; thus, it should be investigated as well.

Conclusions

This study was the first to examine the effects of scale design on ESM measures. Despite previous research suggesting a significant influence on study outcomes, no significant differences between full-labelled scales and endpoint-only-labelled scales could be found on mean scores and measures of emotional variability. Only partly significant effects on distribution measures, the kurtosis of NA measures and the skewness of PA measures, could be found. Thus, one could think about choosing different scale designs for the measures of PA and NA to account for the effects on distribution. As a defining feature of ESM is that the same questionnaire is filled out relatively often, habituation could act as a factor why nearly no differences were found. Despite the efforts of the researchers to have a good study design by having a sufficient sample size and randomization of participants, the fact that most of the participants were students might have influenced the results. For future research, samples should therefore be more diverse. As the effect of scale design on ESM measures has not been investigated before, more research is needed as the possibility that scale design might influence ESM measures could not be ruled out in this study, and other scale designs could be worth investigating.

References

Andrews, F. M., & Crandall, R. (1976). The validity of measures of self-reported well-being. *Social Indicators Research, 3*(1), 1-19. https://doi.org/10.1007/BF00286161

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement, 66*(3), 374-392. https://doi.org/10.1177/0013164405278575

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology, 43*(3), 245-265. https://doi.org/10.1348/0144665031752934

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In *Flow and the Foundations of Positive Psychology* (pp. 35-54). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_3

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity, 52*(4), 1523-1559. https://doi.org/10.1007/s11135-017-0533-4

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting, 27*(2), 35-51. https://doi.org/10.2308/bria-51219

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*(1), 49-57. https://doi.org/10.1177/014662168400800106

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*(2), 176-188. https://doi.org/10.1177/002224379202900203

Gummer, T., & Kunz, T. (2021). Using only numeric labels instead of verbal labels: Stripping rating scales to their bare minimum in web surveys. *Social Science Computer Review, 39*(5), 1003-1029. https://doi.org/10.1177/0894439320951765

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage.

Holmes, E. A., Bonsall, M. B., Hales, S. A., Mitchell, H., Renner, F., Blackwell, S. E., ... & Di Simplicio, M. (2016). Applications of time-series analysis to mood fluctuations in bipolar disorder to promote treatment innovation: a case series. *Translational Psychiatry, 6*(1), e720-e720. https://doi.org/10.1038/tp.2015.207

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin, 141*, 901–930. https://doi.org/10.1037/a0038822

Hoorens, V. (2014). Positivity Bias. In: Michalos, A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_2219

Hurley, J. R. (1998). Timidity as a response style to psychological questionnaires. *The Journal of Psychology, 132*(2), 201-210. https://doi.org/10.1080/00223989809599159

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264-277. https://doi.org/10.1177/0022022104272905

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

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. *Survey Measurement and Process Quality,* 141-164. https://doi.org/10.1002/9781118490013.ch6

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

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education?. *Personality and individual differences, 44*(7), 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales?. *Field Methods, 26*(1), 21-39. https://doi.org/10.1177/1525822X13508270

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology, 44*(1), 369-399. https://doi.org/10.1177/0081175013516114

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry, 17*(2), 123-132. https://doi.org/10.1002/wps.20513

Myin-Germeys, I., & Kuppens, P. (2021). The open handbook of experience sampling methodology. *The Open Handbook of Experience Sampling Methodology; Independently Publisher: Chicago, IL, USA*, 1-311.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Reips, U. D. (2010). Design and formatting in Internet-based research. https://doi.org/10.1037/12076-003

Rodgers, W. L., Andrews, F. M., & Regula Herzog, A. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics-Stockholm-, 8*, 251-251.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582. https://doi.org /10.1086/269282

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in web surveys. *International Journal of Internet Science, 1*(1), 6-16. https://digitalcommons.unl.edu/sociologyfacpub/673/

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116-131. https://doi.org/10.1509/jmkr.45.1.116

Thompson, R. J., Boden, M. T., & Gotlib, I. H. (2017). Emotional variability and clarity in depression and social anxiety. *Cognition and Emotion, 31*(1), 98-108. https://doi.org/10.1080/02699931.2015.1084908

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*(1), 91-112. https://doi.org/10.1093/poq/nfl046

Van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR), 50*(6), 1-40. https://doi.org/10.1145/3123988

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2019). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies, 125*, 118-128. https://doi.org/10.1016/j.ijhcs.2018.12.002

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., ... & Delespaul, P. (2017). The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depression and Anxiety, 34*(6), 481-493. https://doi.org/10.1002/da.22647

van Zutphen, L., Siep, N., Jacob, G. A., Goebel, R., & Arntz, A. (2015). Emotional sensitivity, emotion regulation and impulsivity in borderline personality disorder: a critical review of fMRI studies. *Neuroscience & Biobehavioral Reviews, 51*, 64-76. https://doi.org/10.1016/j.neubiorev.2015.01.001

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. https://doi.org/10.1037/0022-3514.54.6.1063

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236-247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. https://doi.org/10.1177/0013164404268674

Xu, S., & Wang, J. (2019). Still waters stay put: uncovering the effects of emotional variability using experience sampling methodology. *Scandinavian Journal of Hospitality and Tourism, 19*(3), 317-332. https://doi.org/10.1080/15022250.2019.1583124

Appendix A

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:

Name Researcher	***E-mail***
Laura Suntrup	
Lorena Haase	
Gizem Elizabeth Konucu	
Name Supervisor	***Email***
Jannis Kraiss	
Thomas Vaessen	

Appendix B
E-mail and Participation Information

Hello everyone,
Thank you for participating in our study!

Please read the following information carefully **by Sunday**.
Make sure to use the registration code or click on the link below to join the study on Ethica!
It is important that you register by **Sunday at 22:00**.

Briefing

1. General description of the experience sampling method (ESM)

What is ESM? You might not have heard of the term before. ESM is short for Experience sampling Method. This method is used to find out more about daily experiences such as how you feel, which activities you engage in and which people you meet during your day. Your role in this study will be to fill out ten short questionnaires at different times throughout your day for one week. The questionnaires only take about one minute to fill out, so they should not interrupt your daily activities too much. We want to measure life as it is, so it's important that you live your normal life and not adjust your activities to this study. When you hear a beep, open your screen and fill out the short questionnaire immediately. If you only fill it out when you are on your own in a quiet environment, it won't be informative as we are also interested in how you are doing when you are with others, when you are busy, when you are working etc. However, it's okay if on some occasions you happen to miss a beep, we do not want you to adjust your life to the beep. That said, it is of course also important that you do not put yourself into a dangerous situation, such as filling out a questionnaire when you are driving or riding a bike. In general, try to fill it out as quickly as possible after the beep.
Please fill it out on as many occasions as you can.
Do you have any questions?

2. Install Ethica on the participant phone

Download the Ethica app via the play store or app store. Open the app and click on sign up if you don't have an account. If you do, then just log in.
App Store: https://appsto.re/i6h78DQ
Play Store: https://play.google.com/store/apps/details?id=com.ethica.logger
When you don't have an account first click on: You are a participant.
Here you are asked to fill in your first and last name, e-mail address that you would like to use and a password you can easily remember.
When you have done this, a screen will appear where you can enter your registration code for this study.

Your registration code for this study is:
[code]
Or use this link to join the study right away:
https://ethicadata.com/study/2349/

3. Additional information

Contact throughout the study
If you experience any issues with the app throughout the study or are stressed because of the study, do not hesitate to contact one of us, the researchers. Our email addresses and phone

numbers are below this text. We will also send an email on day three of the study to check whether you have any questions.

Lorena Haase
Laura Suntrup
Gizem Elizabeth Konucu

Repeat the most important issues
Now, we will repeat the most important things you need to consider during the study

- Again, it is important that you keep up your normal daily routines. Do not cancel any appointments or make changes to your schedule due to the study
- Try to always carry your smartphone with you.
- Always answer the questionnaire immediately after the beep (of course without creating an unsafe situation).
- Please do not hesitate to get in touch with one of us if you have any issues with the app or experience distress due to the study.
- In case of discomfort and possible psychological distress, you can also contact the health services of the University of Twente. They can help you set up an appointment with one of the student psychologists via email info@campushuisarts.nl or phone +31 53 2030 204.

If you have reached this point, we want to **thank you for informing yourself & again for participating in our study**. We would like to underline one last time that it is of great value that you fill out as many questionnaires as you can!

Kind regards,

Lorena, Gizem and Laura