How to measure how you feel:
Reliability and Construct Validity of the Affect Grid in Experience Sampling

Julia Müller
s1778633
27th June 2019

University of Twente
BMS Faculty
Department of Psychology
Positive Psychology and Technology

Supervisors:
Dr. Matthijs Noordzij
Jannis Kraiss

Abstract

Reliability and validity have not yet been extensively researched in experience sampling methods (ESM). Therefore, this study focused on examining the reliability and convergent validity of the measurement instrument “Affect Grid” in ESM research. The Affect Grid was developed using the Core Affect model and measures two dimensions of feelings, i.e. Arousal and Valence.

Methods

In order to investigate the psychometric properties of this measurement instrument, 19 participants answered six Affect Grids a day, for seven days. Previously, they filled out the neuroticism scale of the short form Eysenck Personality Questionnaire (EPQ-S). To assess reliability, a stability analysis was conducted. Additionally, participants’ neuroticism scores were compared to mean Valence scores, in order to assess the convergent validity.

Results

The stability analysis revealed no fixed effect for time, and correlation analyses revealed significant results for the stability of mean Valence and Arousal scores. In the validity analyses, neuroticism was found to be negatively correlated to Valence. No significant correlation was found between neuroticism and Arousal. However, neuroticism was found to predict both constructs as a covariate.

Discussion

Overall, this research is one of the first on the reliability and validity of the Affect Grid. Besides the limited sample size, the findings suggest that the Affect Grid is a reliable measurement tool. In combination with the evidence for convergent validity, the Affect Grid is recommended to be used in future ESM research. This can be, for example, in domains of therapeutic advancement, prevention of depression or anxiety or in general assessment of disorders involving mood fluctuations.

Table of Contents

1. Introduction

How many times a day do you sit down and really think about what you are feeling? Many people go about their day-to-day lives without spending much time thinking about their feelings. However, feelings and emotions prove to be an important part of the human experience. For example, some individuals that show more emotional variability throughout their lives were found to experience higher levels of neuroticism and depression, as well as low self-esteem (Kuppens, Allen & Sheeber, 2010). Thus, the impact of emotions should not be underestimated. But why should feelings and emotions be measured? Fortunately, this area of research can be applied to many different areas: school and the work field (measuring e.g. stress), psychological research (e.g. therapy and capturing of mental illnesses and their symptoms) and many more. Research on feelings can thus be an important building block in understanding mental disorders, but also a first step towards improving the life and well-being of people in all kinds of living situations.

1.1 Experience Sampling Methods

In the history of Psychology, the focus laid on studying stable traits and feelings for a long time. Over the years, however, there has been a movement towards studying the way in which specific situations, but also contexts, are affecting people's subjective experience of life (Csikszentmihalyi, 2014), especially, as behaviour was found to only be understood within the context that it occurs in (Myin-Germeys et al., 2018). Additionally, it has become known that emotional states, such as happiness for example, are very responsive to and often influenced by the environment. To detect these influences and fluctuations in emotional states, Experience Sampling Methods (ESM) are often used (Csikszentmihalyi & Hunter, 2003). In ESM studies, researchers make use of time-based schemes and, as the name suggests, sample experience. The assessment is focused on repeated subjective self-report measures (Ebner-Priemer, Eid, Kleindienst, Stabenow & Trull, 2009), for example the Affect Grid that will be described later, but also diaries, questionnaires, and other methods. In contrast to classical survey studies, a questionnaire in ESM should not take more than two minutes to be answered, as more items would increase the burden on the participant and by that decrease reactivity to the method (Myin-Germeys et al., 2018).

There are many advantages to ESM, one of them being the ability to capture the variations of emotional states and their contexts within the individual. Its focus is on repeated

measurements throughout contexts and occasions (Csikszentmihalyi, 2014), which allows for proximity to the experience of the participants (Zirkel, Garcia, & Murphy, 2015). In contrast, classical surveys often use measures of self-reflection and ask for the way in which people behave most of the time. This introduces retrospective recall bias, making the answers less expressive (Myin-Germeys et al., 2018). ESM, however, reduces those retrospective recall bias (Versluis et al., 2018) as it investigates the experiences right in the moment, and the individual does not have to think back and explain the context again. Moreover, ecological validity proves to be high, because the assessments occur during real life (Versluis et al., 2018). This recording of answers in the natural environment is a more accurate representation of the behaviour of participants. This is compared to behaviour recorded in the laboratory or in one isolated moment, as it was done earlier in the history of psychology (van Berkel, Ferreira, & Kostakos, 2018).

However, ESM research also faces some problems: first, the researchers ask questions in the middle of someone's life, which is why it is important to consider the time it takes to answer the questions and the number of questions to ask. Too many questions can result in a greater burden on the participant which may in turn result in lower reliability and validity (Versluis et al., 2018). This is because participants may stop answering questions fully or honestly. Second, the cooperation of participants with the study is strongly dependent on the participant's trust and commitment to the research, as well as their belief in the research (Csikszentmihalyi, 2014). Here, it is important to consider that ESM can only use information that participants are willing to share (Christensen, Barrett, Bliss-Moreau, Lebo & Kaschub, 2003). Third, it is often possible that uncontrolled or unobserved variables exist, that would change or even eliminate the associations found when controlled for (Carlson, 2019). Thus, when conducting an individual case study, generalisations cannot be made from one individual to another (Csikszentmihalyi, 2014).

Contrasting to the many, thorough, studies of reliability and validity of classical questionnaires (e.g. the Mental-Health Continuum Short-Form investigated by Lamers, Westerhof, Boglmeijer, ten Klooster, and Keyes in 2010) only a few researchers, including Csikszentmihalyi (2014) and Versluis et al. (2018), have already studied the reliability and validity of ESM research. Csikszentmihalyi (2014) mentions a study that investigated individual differences in people suffering from bulimia. The researchers used a pager to assess hour-to-hour moods and later used split-half reliability to investigate the reliability of the assessment. It was found that "*individual responses are relatively stable*" (Csikszentmihalyi, 2014). This means that people can have different fluctuations in feelings, but those

fluctuations are stable across the measurement period. Versluis et al. (2018) adapted an already existing questionnaire to the ESM method, to enable measuring emotional awareness in everyday life. They found that the ESM method does capture the individual differences meaningfully and validly, but there was a lot of unexplained variance across participants. Therefore, they suggested that more research is needed to understand the psychometric properties of ESM research. Hence, this is the main focus of this paper. Here, in order to assess convergent validity, outcomes of already validated tests will be compared to the outcomes of the present ESM research. In this case, as will be explained later, the neuroticism scores of the short-form revised Eysenck Personality Questionnaire (EPQ-S, Francis, Lewis, & Ziebertz, 2006) will be used to assess the convergent validity of the used ESM measure.

1.2 Core Affect and Affect Grid

To be able to provide research on the reliability and convergent validity of ESM research on emotions and feelings it is first important to define the latter. This, however, may be difficult. There are too many words to describe feelings and emotions to arrive at one final definition. Take sadness as an example: one can feel sad, down, lonely, heartbroken, ... the list is long. Additionally, experiences may vary from person to person. As Richards (2010) states: "*we may try and communicate [experiences] via another medium, like music or painting, but even then, we cannot, by definition, say what has been communicated or know that we have succeeded*". Hence, in this report the concept of "Core Affect" will be used to measure feelings, to avoid countless definitions and confusion with interpretations of emotions and feelings. Core Affect is essentially the most readily accessible affective feelings: pleasure or displeasure, depression or elation, and tension or relaxation (Russel & Barrett, 1999). It describes an integral state of feeling that represents a combination of Arousal and affective Valence, both constructs being independent of each other (Ramos, Sendra, Sánchez, & Mena, 2017). Russel and Barrett (1999) claim that a person always finds him or herself in a state of Core Affect that varies in intensity, even when feeling neutral.

Core Affect can be measured using the Affect Grid (Figure 1). Here, the participant can indicate his feelings in terms of Arousal and Valence in a grid. The scores on each concept can reach from one to nine (in this example), one indicating low Arousal or unpleasant feelings, nine indicating high Arousal and pleasant feelings (Russell, Weiss, & Mendelsohn, 1989). This method is easy to use and compared to other measures affords less detail (Russel & Gobet, 2012). Unfortunately, the Affect Grid appears to be slightly less reliable than other self-report questionnaires measuring mood when used as a one-question variant (Russell,

Weiss, & Mendelsohn, 1989), which is why it is important to investigate its reliability in ESM studies, when used repetitively, in more detail.

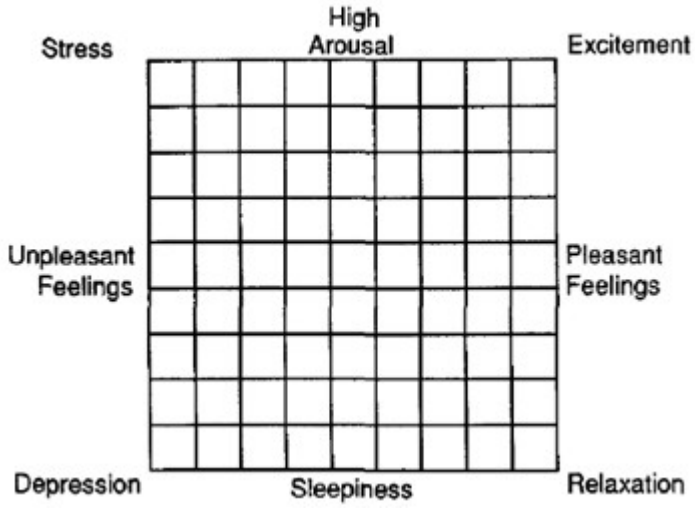


Figure 1. Affect Grid developed by Russel, Weiss & Mendelsohn (1989)

1.3 Neuroticism

Neuroticism is a personality trait belonging to the so-called Big Five (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) and is thus an important source for individual differences. Moreover, the Big Five are most the widely studied personality factors as they consistently predict life outcomes meaningful to individuals. People that score high on neuroticism are known to respond negatively and disproportionately to challenges that face them from day to day (Miranda & Blais-Rochette, 2018). Furthermore, years of research have shown that individuals scoring high on neuroticism more often suffer from episodes of negative emotions such as anxiety or distress (Rusting & Larsen, 1997). It is therefore no surprise that neuroticism has been found to be linked to anxiety and depressive symptoms, as well as mood instability, which refers to having extreme and frequent fluctuations of mood over time (Bowen, Baetz, Leuschen, & Kalynchuk, 2011). Kuppens, Oravecz, and Tuerlinckx (2010) mention that neuroticism is regarded as a predictor and indicator of lower levels of Valence, indicating lower moods. Valence, like neuroticism, was found to have a negative relationship with depression, general disturbance, and suicidality (Ramos, Sendra, Sánchez, & Mena, 2017). In contrast, Arousal has not been found to be related to neuroticism in previous studies, as it promotes emotions such as rage or excitement (Ramos, Sendra, Sánchez, & Mena, 2017), while individuals

scoring high on neuroticism are characterised by feelings of tension or worthlessness (Cummins, Stokes, & Davern, 2007).

In conclusion, it can be said that a combination of high neuroticism and low Valence scores can lead to symptoms of psychopathology, such as anxiety or depression. This is why it becomes important to investigate the relationship of the two constructs in the average answers of the participants on the Affect Grid. Therefore, scores on neuroticism will be compared with the participants' average scores of Valence in the Affect Grid. Average Arousal scores will be tested for correlation as well, to check the accuracy of the assumption that there is no relation between the two constructs. The focus of this paper will lay on the average scores of the Affect Grids, following the suggestion of Myin-Germeys and colleagues (2018) to investigate the average levels of negative affect within a group of participants. In this case, the aim is to find out whether participants scoring high in neuroticism show lower average levels of Valence, than participants scoring low in neuroticism.

1.4 Research Questions

Overall, the focus of this study is on the reliability and convergent validity of the Affect Grid as a repeated measurement tool in an ESM study. As already mentioned, there has been no research yet about the psychometric properties of the Affect Grid used repeatedly and thus, it is important for the research field of ESM to investigate these properties. The reliability will be assessed by conducting a stability measure, comparing participants' average scores of the first half of scores on the Affect Grids to the second half. It is expected that the Affect Grid measures Arousal and Valence in a stable manner over time. The first research hypothesis is therefore that the means of the Affect Grid are stable over time. To assess convergent validity, scores on neuroticism and the average scores on Valence of the participants' answers will be compared. In this research, it is expected that when a participant's score on neuroticism is high, lower average scores on Valence are present. The second hypothesis being that the scores on neuroticism predict the average Valence scores. Additionally, it will be controlled for a relation between Arousal and neuroticism. It is expected, that no relation between these two constructs will be found, which is the third hypothesis of this study.

2. Methods

2.1 Design

The present one-group survey design experience sampling study was developed to evaluate the reliability and convergent validity of an Affect Grid used in ESM, where fluctuations within the participant can be investigated. First, the participant was asked to fill out three questionnaires (explained in the “materials”). The EPQ-S will help to investigate the convergent validity of the Affect Grid in the data analysis. The Affect Grid was implemented using a smartphone app developed by researchers at the University of Twente (The Incredible Intervention Machine (TIIM)). It was available for participants to answer six times a day, for seven days. Thus, participants were asked to respond 42 times. All in all, there were therefore 4 measurements: three questionnaires and, repeatedly, the Affect Grid. In this study, the researchers made use of signal-contingent reporting, meaning that the participants received a notification on their device when there were questions available to be answered.

2.2 Participants

For the present ESM study, participants were asked to participate using the convenience sampling method, meaning that the researchers asked people in their environment via text messages and social connections to participate. In total, there were N=52 participants that agreed to engage in the study. Of this, 34 (65.4%) were female and 18 (34.6%) male, with an average age of 22.6 years ranging from 18 to 52 (SD=5.8). 50 of the participants were German, two of other nationalities. Already existing studies were used to guide the decision of setting the compliance rate. Chen, Cordier, and Brown (2015) suggest, that in ESM research it is commonly advised to only use participants’ scores with a compliance rate above 33%, which would, in this case, equal to 15 responses. However, Palmier-Claus and colleagues (2011) recommend using those participants’ scores that answered at least 20 of the items. Thus, before the implementation of the study it was agreed to only use the data of those participants that answered at least 28 (=65%) of the Affect Grids, finally excluding 26 people from the data analysis. Additionally, technical problems with the smartphone application appeared for 7 people, making them unable to follow the intervention. Overall, 33 participants (63.5%) were excluded from the study, leaving an overall sample of N=19 for the data analysis. Of these, 15 participants scored an additional day of Affect Grids, resulting in more than 42 answers for some of them. The data of the additional day, however, will only be included for the descriptive statistics, as they do not fit with the stability

measures for reliability. Furthermore, due to technical error, another four participants were not shown the neuroticism scale of the EPQ-S and were therefore unable to answer it, which excludes them from the correlation analysis between neuroticism scores and the average scores on Arousal and Valence, leaving a sample of N=15 for the validity analysis. .

2.3 Materials

Participants were asked to fill out the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988) the neuroticism scale of the EPQ-S (Francis, Lewis, & Ziebertz, 2006) and the TAS-20 (Bagby, Taylor, & Parker, 1994), while in the present report only the neuroticism scale by Eysenck will be considered throughout the data analysis and further interpretations. Additionally, screening questions about the age, gender and nationality of participants were asked. In the following, the application TIIM, the neuroticism scale of the EPQ-S, and the Affect Grid will be described in further detail.

The Incredible Intervention Machine (TIIM). This survey smartphone application was developed by researchers at the University of Twente. It is available for download for IOS and Android and helps researchers to implement an intervention within a group of participants ("TIIM (The Incredible Intervention Machine)", 2018). In the way of ESM, questions are made available to the participant on set points in time, and participants receive a notification on their smartphone when the question is available. Additionally, the questions expire after a set time frame.

Researchers have the opportunity to customize the intervention for their research purposes, making use of different measurement tools (e.g. Likert Scales, free text, or, like in this research, an Affect Grid, …). Moreover, the time slots can be individualised as well. The participants register for the intervention on any device via a website that works as a starting page for the research. Here, screening questions such as gender or age can already be asked. After registering, the smartphone app has to be downloaded from the App/Play Store for the participant to be able to start the intervention on their smartphone.

Neuroticism-Scale. A module was established within TIIM, adding 12 neuroticism Items of the EPQ-S (see appendix A) to be answered on a 5-point Likert Scale. For example, one question being "Are you an irritable person?" with the answer options "not at all, a little, moderately, a lot, very much". Muñiz, García-Cueto, and Lozano (2005) mention that many studies show that Likert-type question format improves reliability and validity, as well as

providing a higher understanding of the person taking the test and a clearer structure of the dimensions of the questionnaire. Additionally, the Likert-type questions allow for greater flexibility, not forcing the participant to choose from two answers, raising the satisfaction of participants (Muñiz, García-Cueto, & Lozano, 2005). Specifically, for the EPQ-S the above-mentioned authors found that the questionnaire still shows good psychometric properties when replacing the original dichotomous answer options with Likert-Scales. As people that score high on neuroticism tend to fluctuate more in their moods, using the participant's score on neuroticism and comparing it to their overall fluctuations in the Affect Grid will help to control for convergent validity.

Affect Grid. As previously mentioned, the Affect Grid (Figure 1) was developed by Russell, Weiss, and Mendelsohn, in 1989 to assess the Core Affect of people, measuring Arousal and Valence. For the present study, the Affect Grid had to be adjusted to fit the boundaries of the TIIM image display and underlying axis system. For this, the labels of axes need to be within the figure. Additionally, the original labels "Depression, Relaxation, Excitement, and Stress" were not included in the figure, as these might result in so-called "lexical priming" and elicit a negative evaluation of the words and associated feelings (Hoey, 2012). This might then result in different answers and therefore it was decided to not include the words. The edited Affect Grid implemented in the TIIM can be seen below in Figure 2. Another thing that is important to mention is that the measuring scale of the TIIM differs from the Affect Grid, as the Affect Grid shows a 10x10 design, while TIIM measures from -100 to 100. However, the TIIM's scale was found to be too vague for participants to answer, as there is no way to indicate a difference between feeling a 36 in Arousal, or a 37. Thus, it was agreed to use the 10x10 Grid for the visual design but use the TIIM's scale in the analyses.

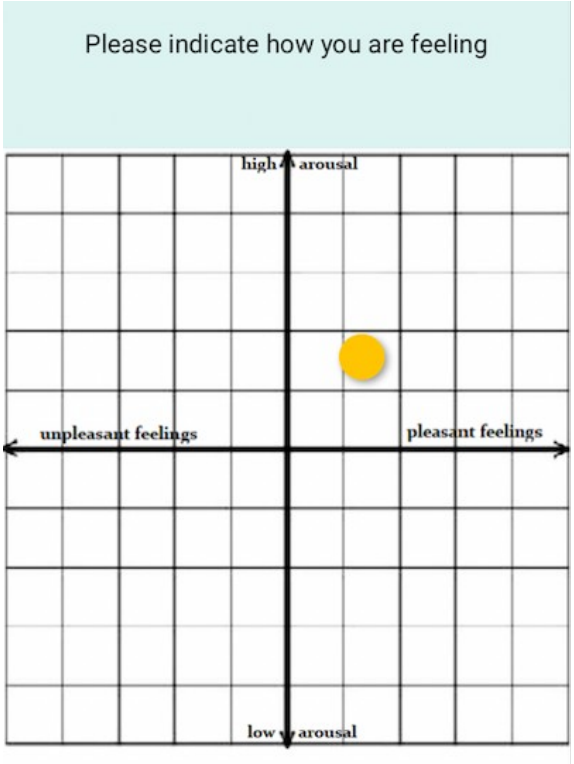


Figure 2. Adjusted Affect Grid in the smartphone application TIIM. Participants were able to move the yellow dot across the Grid, indicating how they were feeling at the specific moment in time.

2.4 Procedure

Before implementing the study and sending it to possible participants, the researchers pilot-tested the study on themselves. A one-day pilot was established, where the researchers answered the three afore-mentioned questionnaires and six Affect Grids. This enabled recognition of possible problems and made possible to fix them. For example, in the Pilot, problems with the time schedule of the Affect Grids became apparent. Every Affect Grid was shown from the beginning, while they were supposed to be shown every two hours. Thus, the time schedule was adapted and improved, making it possible to implement the intervention.

After the study was approved by the Ethics Review Committee of the University of Twente in March 2019, the data collection was started in April 2019. The researchers sent the link to the starting page of the research to their social environment via WhatsApp and other Social Media. The prospective participant was then able sign up with an e-mail address and password. He had to sign the informed consent and answer three screening questions, asking for gender, age, and nationality, to proceed. After that, he was added to the participant list in TIIM and asked to download the app to his smartphone and to log-in to the system.

The researchers were then able to assign the participant to an intervention established within TIIM, starting on the next day at 10 am, where an e-mail was sent to the participant explaining the procedure and the Affect Grid (see appendix B). Interventions were started on three separate dates, as participants were asked to participate at different points in time (05.04.19, 07.04.2019, and 12.04.2019). Once started, the participant was asked to fill out the three mentioned questionnaires (i.e. the neuroticism scale, the PANAS, and the TAS-20). This is equal to 50 items, taking about 15 minutes to answer. After that, he was able to fill out the first Affect Grid. Whenever a new Affect Grid was supposed to be answered (every two hours from 10am to 8pm) the participant received a notification on his smartphone. He was then able to answer the Affect Grid for 90 minutes, before it disappeared from the app. Answering the Affect Grid only took about a minute of his time, as he only needed to set a point in the coordinate system that indicates the current state of feeling. There were six Affect Grids to be answered on a day, for seven days a week, equalling 42 Affect Grids. Taken together, this study, including the 42 Affect Grids, the three questionnaires and the sign-up procedure, took about an hour of the time of the participants spread across one week.

2.5 Data Analysis

Prior to the data analysis, the data was imported from the app TIIM into SPSS (version 25). The data set was then adjusted to fit the analyses. Thus, the data set was edited into a wide format. Additionally, Arousal and Valence were separated into two variables, as they are found to be two different and independent constructs in emotional experiences (Ramos, Sendra, Sánchez, & Mena, 2017). Then, general descriptives were calculated to get an overview of the data. Additionally, two boxplots were calculated to get an idea of the general Arousal and Valence scores of participants. Next, the average score on neuroticism was calculated to then compare individual participant’s scores on neuroticism with their mood fluctuations in the Affect Grid as seen in the boxplots. As the answer options on the neuroticism scale ranged from one to five, one (“a little”) indicates a low score on neuroticism and five (“a lot”) indicates a high score on neuroticism. To assess the reliability of the Affect Grid, an analysis of stability was conducted between the mean of the first 21 answered Affect Grids, and the remaining 21, as well as the number of even and odd Affect Grids. Here, using both forms of splitting up the data, the meaning of the conclusions about the reliability of the Affect Grid is increased. To reach a conclusion, scatterplots were established. After that, the correlation coefficients between each of the two halves were calculated. Additionally, to

assess the convergent validity of the Affect Grid, a linear mixed model analysis was conducted. Here, it was determined whether time has an effect on the scores of the Affect Grid. Also, the model yielded overall means for Arousal and Valence for each person, enabling the calculation of the correlation coefficients of each concept with neuroticism. Afterwards, it was tested for fixed effects between neuroticism and the two Core Affect constructs. Additionally, with these scores of the linear mixed model analysis, a line diagram was established to illustrate the effects that were found. In general, throughout the data analysis, correlations are regarded as very strong when >0.8/-0.8, as strong between 0.6-0.79, as moderate between 0.59- 0.4, and as weak when lower than 0.39 (Swinscow & Campbell, 2002).

3. Results

3.1 Descriptive Statistics

Affect Grid. To get an overview of the scores of the sample, the mean of all scored Affect Grids was calculated. Overall, 795 Affect Grids were answered by the 19 participants included in this analysis. The mean score for Valence (M=33.31) was higher in the overall sample than that for Arousal (M=14.39). When looking at individual scores, in the boxplots (Figure 3 & 4), this tendency of higher mean Valence scores than mean Arousal scores is apparent as well. Each individual's mean scores can be found in Appendix C.

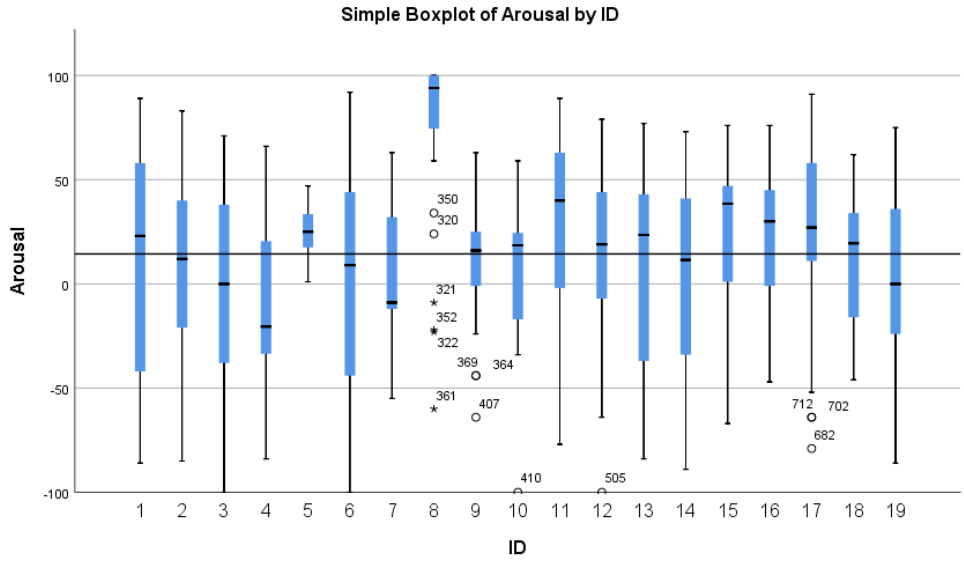


Figure 3. Boxplot of Arousal scores per individual, including mean (M = 14.39)

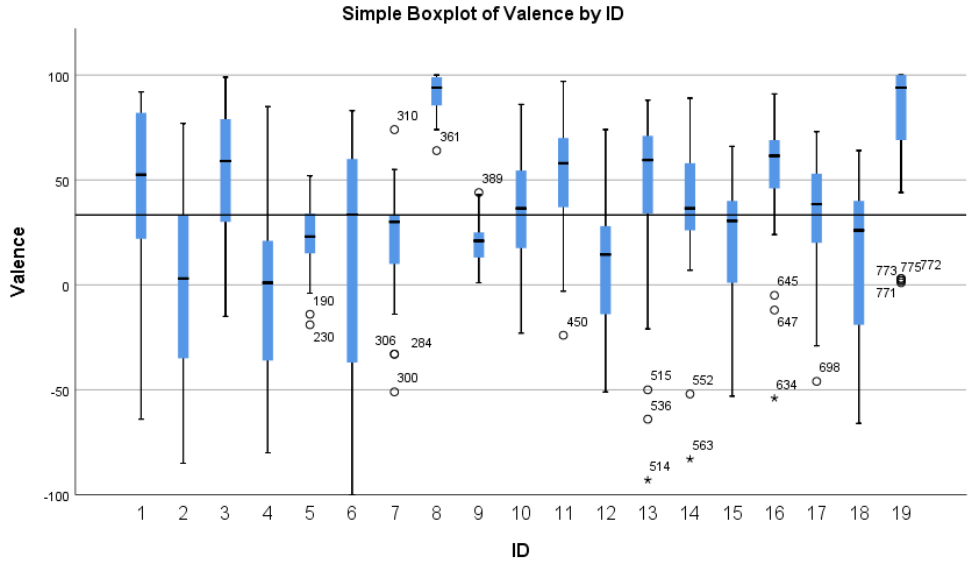
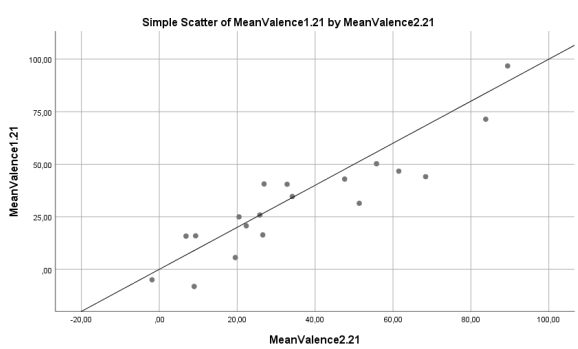


Figure 4. Boxplot of Valence scores per individual, including mean (M = 33.31)

Neuroticism Scale (EPQ-S). In general, the sample scored a 2.69 on the neuroticism Scale, meaning that, overall, the sample scores neither high, nor low on neuroticism. Every participant's mean score on neuroticism can be found in Appendix D. On the individual level, only three participants scored lower than two (1.33, 1.33, and 1.83), therefore scoring low on neuroticism. When compared to their average scores on Arousal and Valence, participant number 8 (with a low score of 1.33 on neuroticism), for example, also shows little variance in the scores on Valence and Arousal, indicating little mood fluctuations. Opposed to this, participant number 1, scoring the highest on neuroticism (3.58) shows a lot of variance across Valence and Arousal scores, indicating higher mood fluctuations. The mood fluctuations are visualised in the boxplots (Figure 3 & 4).

3.2 Reliability Assessment

Stability Measures. As can be seen in the scatterplots for Valence scores, a linear relationship between the first and second halves, as well as even and odd Affect Grids is apparent (Figure 5a and b). The Pearson r correlations between the even and odd numbers and the first half and second halves were found to be very strong ($r=.96$; $p<.01$ and $r=.91$; $p<.01$, respectively).



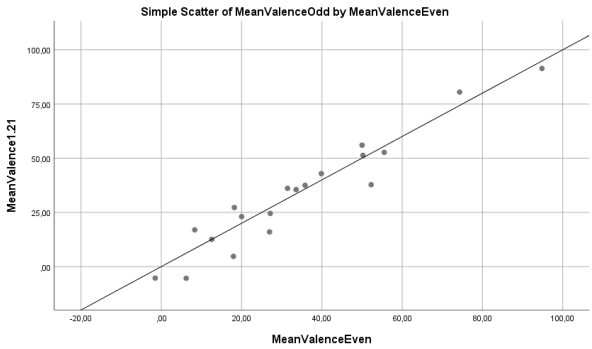
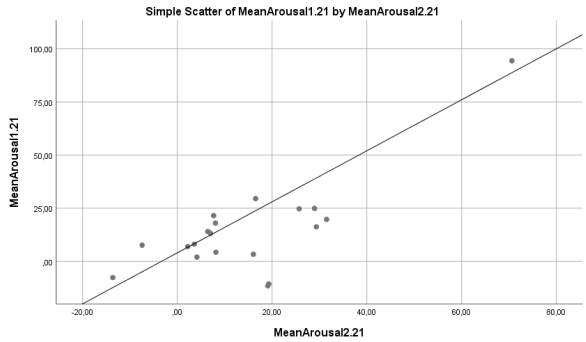


Figure 5a and 5b. Scatterplots showing the relationship between the first and second half of answers as well as even and odd numbers of answers indicating Valence. A linear relationship can be assumed from these graphs.

For Arousal, however, no linear relationship can be found in the scatterplots between both first and second halves as well as even and odd numbers of the Affect Grids (Figure 6a & b). Both Spearman correlation analyses for the mean scores of Arousal on the first and second half, as well as even and odd numbers of Affect Grids were found to be moderately significant (rho=0.47, p=.04; rho=.46, p=.04, respectively).



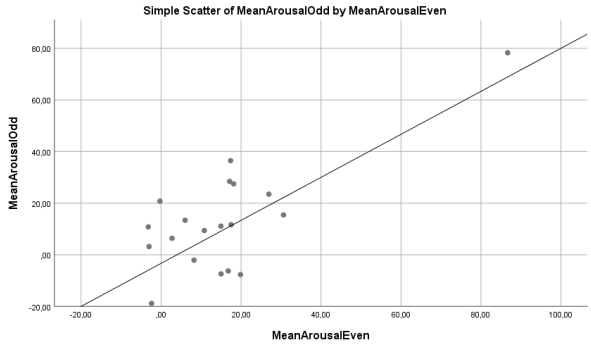


Figure 6a and b. Scatterplots showing the relationship between the first and second half of answers as well as even and odd numbers of answers indicating Arousal. No linear relationship can be assumed from these graphs.

Additionally, it was important to check whether over time there was a linear growth in the answers to the Affect Grid. As expected, it was found that there was no significant effect of time on the responses to the Affect Grid in Arousal (F=1.20; p=.18) or Valence (F=1.04; p=.40).

3.3 Validity Analysis

To assess convergent validity, a correlation analysis of the estimated marginal means of Valence resulting from linear mixed model analyses and the means of neuroticism was conducted. For Valence, a weak linear relationship can be assumed from the distribution of points in the scatterplot (Figure 7). Valence and neuroticism are shown to correlate significantly (r = -.62; p = .01), showing a negative relationship between Valence and neuroticism scores. Additionally, when testing for fixed effects between the dependent variable of Valence and the neuroticism scores, neuroticism was found to be a significant covariate (df = 102.87; F = 41.37; sig = .00). This means that the neuroticism score can predict the mean Valence score, thus indicating significant convergent validity. To illustrate the relationship found between neuroticism and Valence as measured in the Affect Grid, a line diagram was established (Figure 8). Here, it can be seen that indeed lower neuroticism scores may indicate higher scores on Valence (e.g. participants 8 and 15) and that higher neuroticism scores result in lower Valence scores (e.g. participants 1 and 4).

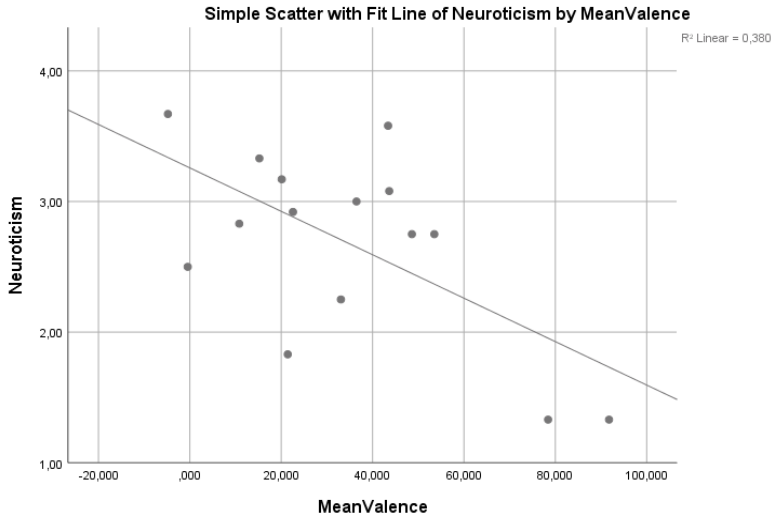


Figure 7. Scatterplot of the relationship between mean scores on Valence and neuroticism, indicating a weak negative linear relationship.

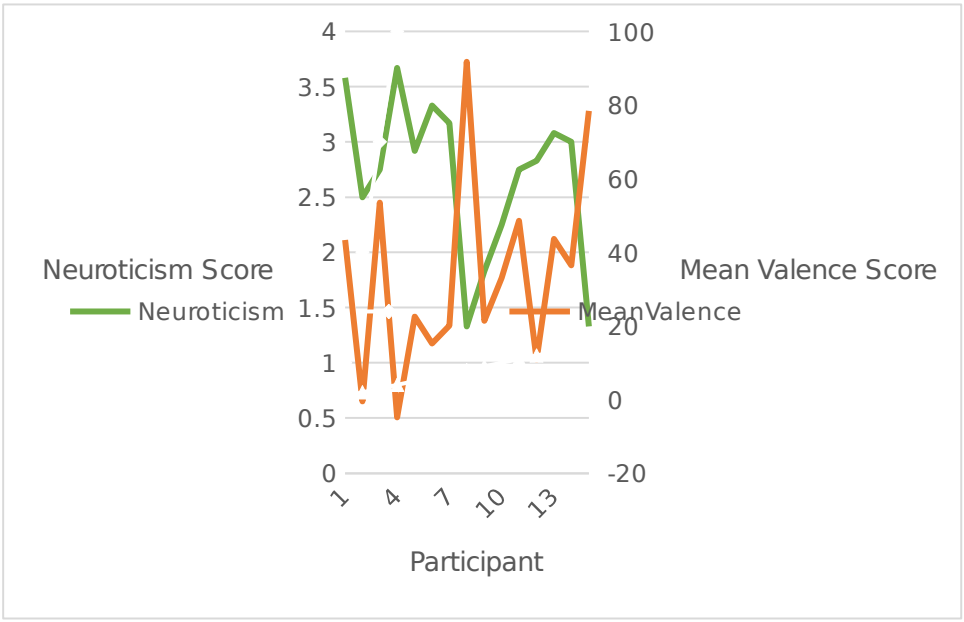


Figure 8. Line diagram of the mean scores of Valence and neuroticism per participant.

When testing for a relationship between Arousal and neuroticism, as expected, no linear relationship could be assumed from the scatterplot (Figure 8). Arousal and neuroticism do not correlate significantly (rho= -.49; p = .06), showing that there is no relationship between both constructs. However, in contrast to this finding, the testing for fixed effects between Arousal and neuroticism indicated that mean Arousal is also significantly predicted by the neuroticism score of the participant (df = 136.74; F = 23.68; sig. = .00).

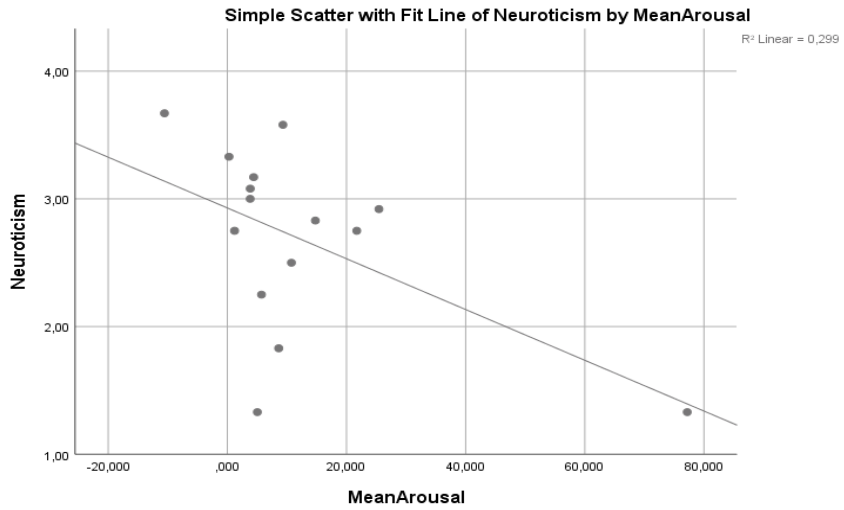


Figure 9. Scatterplot of the relationship between mean scores on Arousal and neuroticism. No linear relationship can be assumed from this graph.

4. Discussion

Based on the Core Affect model described by Russel in 1999, this study's purpose was to test for the reliability of the Affect Grid measuring Arousal and Valence in ESM research, as well as for the convergent validity. When testing for stability, it was found that the Affect Grid is a reliable measure of Arousal and Valence. By comparing the resulting mean Valence scores of the Affect Grid to scores of the already existing neuroticism scale of the EPQ-S, it was found that the Affect Grid is a valid measure of Valence in ESM research. In the following, more attention will be brought to the conclusions that can be drawn from the results. That is, considering the relationship between neuroticism and both constructs of the Core Affect and considering the Affect Grid in ESM research. Additionally, limitations will be discussed. Finally, some suggestions for the integration of the Affect Grid into future research will be provided.

The first hypothesis ("The means of the Affect Grid are stable over time") was accepted after testing for stability among the first and second half of answers, as well as even and odd numbers of answers. Here, for the construct of Valence, very strong evidence for reliability was found. For Arousal, moderate evidence for reliability was found, as the correlation analysis yielded moderately strong correlation coefficients. The findings are in line with what was found by other researchers before. For example, in a study of individual differences in bulimic individuals, Csikszentmihalyi (2014) describes that these differences are relatively stable in the ESM study. This means that the results remain stable across several

measurements points. This was also found in the present study: there was no significant effect of time on the Affect Grid, indicating that there is no linear growth or decrease in the data. The participants' mean scores are therefore stable across the measurement period. As Russel, Weiss, and Mendelsohn (1989) mentioned, the Affect Grid appears less reliable than other self-report questionnaires that measure mood when used as a one-question variant. The present study's results suggest accepting the first hypothesis, indicating that the reliability of the Affect Grid is strong when used as a repeated measurement instrument in ESM research.

In 2018, Myin-Germeys and colleagues (2018) suggested, that there is increased need for "*robust psychometric research for the structural validity of ESM measures of key constructs*…". Thus, testing the second hypothesis ("Scores of neuroticism predict the average Valence scores of Participants") was of particular importance for testing the convergent validity. From the results it became apparent that neuroticism is a significant covariate for Valence, showing that it does predict the average scores. These results therefore suggest accepting the second hypothesis. Additionally, low Valence was frequently found to be related to neuroticism (Kuppens, Oravecz, & Tuerlinckx, 2010; Rusting & Larsen, 1997). As low Valence may result in depression and suicidality (Ramos, Sendra, Sánchez, & Mena, 2017), which are both frequent consequences of high neuroticism (Bowen, Baetz, Leuschen, & Kalynchuk, 2011), the correlation between neuroticism and Valence is understandable. The significant negative correlation also indicates that these two constructs are related. The higher the score on neuroticism, the lower the scores on pleasurable feelings, i.e. Valence. This might also be the other way around; lower Valence scores might indicate higher scores on neuroticism.

Underlining this finding, it is known that individuals that score high on neuroticism show episodes of negative emotions such as anxiety or distress, much more often than individuals scoring low in neuroticism (Rusting & Larsen, 1997). Additionally, Kuppens, Oravecz, and Tuerlinckx (2010) found in their study that neuroticism was related to more unpleasant average scores for the individual. The same was found when testing the theoretical models underlying this assumption. Rusting and Larsen (1997) tested whether and in which way Extraversion and neuroticism were related to positive and negative affect and found that indeed, neuroticism is related to negative affect, the increased feeling of negative emotions. Overall, the findings indicate high convergent validity of the Affect Grid, as neuroticism has been shown to strongly predict the mean scores on Valence, accepting the second hypothesis.

The third hypothesis ("There is no relation between neuroticism and Arousal") was confirmed by the present results, as Arousal and neuroticism did not show evidence for a

linear relationship in the scatterplot. Additionally, they did not correlate significantly with each other. Evidently, Arousal was found to be related to impulsivity and a tendency towards action (Ramos, Sendra, Sánchez, & Mena, 2017) while high neuroticism often relates to low moods and less activity among individuals (Bowen, Baetz, Leuschen, & Kalynchuk, 2011). It is not surprising that these two constructs do not significantly correlate with each other. In other research, it was also found that there is no relationship between neuroticism and volatile Arousal, meaning that, when tested for over different points in time, neuroticism and Arousal do not relate to each other (Kuppens, Oravecz, & Tuerlinckx, 2010).

However, this research yielded mixed evidence, since when testing for neuroticism as a covariate of Arousal, weak, but significant effects were found. This does indicate that neuroticism predicts mean Arousal scores. In contrast, no significant relationship was found between mean Arousal and mean neuroticism scores in the correlation analyses. A study by Kuppens et al. (2017) may help to explain this mixed evidence. In this study, it was found that individuals with high scores on neuroticism felt high Arousal combined with negative Valence as well as negative Valence combined with low Arousal. Thus, neuroticism can result in experiencing both high and low Arousal but remains stable with experiencing negative Valence, which underlines the existence of the relationship between neuroticism and Valence and explains the lack thereof between neuroticism and Arousal. Additionally, Kuppens and colleagues' (2017) finding underlines the previously mentioned finding of this study. Namely, that neuroticism and Valence show a strong relationship.

4.1 Limitations

A major strength of the current study presents itself in the fact that not much research has yet been conducted in the domain of psychometric properties of the Affect Grid in ESM research. However, as any study, this research also faced limitations. The research itself presents to be very innovative, however a generic limitation has to be mentioned. Unfortunately, the sample size for the analysis was very small. This reduces the power of the statistical analysis and can result in lower validity of the whole study (Versluis et al., 2018). Additionally, this raises questions about the outcomes of this study. With more participants, the outcomes may be different. The reliability of the Affect Grid may increase or even decrease. The same may happen with convergent validity. As Faber and Fonseca (2014) explained, a proposed measurement tool might have no disadvantages when compared to traditional instruments. The found inferiority might result from a small sample size. The same can happen here: The Affect Grid might only show evidence of reliability and validity because

of the sample size. Next, this study was only conducted over the course of a week. If the measurement period is extended, the results could change. As Versluis and colleagues (2018) mentioned, a greater burden put on the participant can decrease the reliability of the study. Thus, when used repeatedly for a longer time, the reliability estimation of the Affect Grid could decrease, contrasting the present findings.

Additionally, in this research only correlational analyses were conducted. From the results one can assume a relationship between two variables, but the correlation does not equal to causation. Thus, low neuroticism scores may not cause low Valence scores, or the other way around. From the research results it can therefore only be assumed that there is a negative relationship between the two constructs. However, this can result from other, not controlled for variables (Carlson, 2019). Moreover, in this study's sample, no high mean Arousal scores were found. In a population with higher Arousal scores, the results of the correlation analysis may turn out different, and the mixed evidence of the relationship between neuroticism and Arousal could become clearer. This means that there could be increased evidence for no relationship between both constructs. This can in turn further increase the strength of the relationship that was found between neuroticism and Valence.

Lastly, another limitation could be seen in the fact that peoples' daily lives are interrupted by the notifications and answering to the Affect Grid, which might make answers less honest or thought through. However, this is essentially the goal of ESM research. Its aim is to investigate the feelings, emotions, and behaviours of people right in the moment (van Berkel, Ferreira, & Kostakos, 2018). Thus, it is unavoidable to intervene in the daily lives of the participants with the ESM method.

4.2 Implications and Future Studies

This study essentially focused on the mean scores of participants and its implications. In future research the focus could, however, be shifted towards the variability of scores. In general, the variability of individual scores was found to provide a method to quantify dynamic processes of individuals (Du & Wang, 2018). Within the context of feelings and emotions, the variation and fluctuations are important in many mental disorders, and ESM has been shown to measure these fluctuations in more detail (Myin-Germeys et al., 2018). With the Affect Grid, fluctuations of feelings can be compared to the mental health state of participants. Du and Wang (2018) found that the number of measurement occasions is very important to increase the reliability of indicators of variability in individuals. Thus, coming back to the fact that this study was only conducted over the course of one week, the reliability

of the Affect Grid could increase when measuring the variability of feelings for a longer period of time. Additionally, a longer time of measurement helps to explore the relationships between indicators and variables (Du & Wang, 2018), in this case it could help to further explore the relationship between neuroticism and valence.

As the mean Valence scores have shown to be predicted by neuroticism scores, the results of future studies like this one can improve the understanding of for example depression and anxiety. The Affect Grid can assumedly reliably yield measures of when and how often participants feel low Valence and Arousal, indicating lower moods. This can advance the research on mood fluctuations in individuals suffering from mental disorder. Additionally, as also found by Myin-Germeys and colleagues (2018), it can advance therapeutic progress. Furthermore, it might be helpful in preventing the development of a disorder in individuals that are prone to have higher scores of neuroticism and lower scores of Valence, when combining the Affect Grid with an intervention.

4.3 Conclusion

As there has been a need for the investigation of the psychometric properties of ESM studies (Myin-Germeys et al., 2018), this study slightly advances the research field. Not many researchers have yet tested for the reliability and validity of ESM research. Those who did, for example Versluis and colleagues (2018), found that ESM is a reliable way of measuring emotions. However, they also suggested to provide more research in the future. The present paper describes one of the first methodologically sound research on the reliability and validity of the Affect Grid in ESM research. Besides the limited sample, the results of the research give incentives to future research to use the Affect Grid in domains where feelings need to be measured repetitively. The Affect Grid was found to be strongly to moderately reliable. Additionally, it is assumed to have strong convergent validity, as the relationship between Valence and neuroticism was shown to be significantly strong. In conclusion, this research study shows a first indication that in an ESM study, the Affect Grid can reliably and validly be used.

References

Bagby, R. M., Taylor, G. J., & Parker, J. D. (1994). The twenty-item Toronto Alexithymia Scale—II. Convergent, discriminant, and concurrent validity. *Journal of psychosomatic research*, *38*(1), 33-40.

Bowen, R., Baetz, M., Leuschen, C., & Kalynchuk, L. E. (2011). Predictors of suicidal thoughts: mood instability versus neuroticism. *Personality and Individual Differences*, *51*(8), 1034-1038.

Carlson, B. W. (2019, January 07). Simpson's paradox. Retrieved May 16, 2019, from https://www.britannica.com/topic/Simpsons-paradox

Chen, Y. W., Cordier, R., & Brown, N. (2015). A preliminary study on the reliability and validity of using experience sampling method in children with autism spectrum disorders. *Developmental Neurorehabilitation*, *18*(6), 383-389.

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*(1), 53-78.

Csikszentmihalyi, M. (2014). *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi*. Berlin: Springer. doi:10.1007/978-94-017-9088-8

Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. *Journal of happiness studies*, *4*(2), 185-199.

Cummins, R. A., Stokes, M. A., & Davern, M. T. (2007). Core affect and subjective wellbeing: a rebuttal to Moum and Land. *Journal of Happiness Studies*, *8*(4), 457-466.

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T.J. (2009.) Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology, 118*(1), 195-202. https://doi.org/10.1037/a0014868

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental press journal of orthodontics*, *19*(4), 27-29.

Francis, L. J., Lewis, C. A., & Ziebertz, H. (2006). The short-form revised Eysenck Personality Questionnaire (EPQ-S): A German edition.

Hoey, M. (2012). *Lexical priming: A new theory of words and language*. Routledge.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological science*, *21*(7), 984-991.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042-60. doi:10.1037/a0020962

Kuppens, P., Tuerlinckx, F., Yik, M., Koval, P., Coosemans, J., Zeng, K. J., & Russell, J. A. (2017). The relation between valence and arousal in subjective experience varies with personality and culture. *Journal of personality*, *85*(4), 530-542.

Lamers, S. M., Westerhof, G. J., Bohlmeijer, E. T., ten Klooster, P. M., & Keyes, C. L. (2011). Evaluating the psychometric properties of the mental health continuum-short form (MHC-SF). *Journal of clinical psychology*, *67*(1), 99-110.

Miranda, D., & Blais-Rochette, C. (2018). Neuroticism and emotion regulation through music listening: A meta-analysis. *Musicae Scientiae, (201810)*. doi:10.1177/1029864918806341

Muñiz, J., García-Cueto, E., & Lozano, L. M. (2005). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and individual differences*, *38*(1), 61-69.

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: Official Journal of the World Psychiatric Association (wpa),17*(2), 123-132. doi:10.1002/wps.20513

Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., ... & Dunn, G. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12-20.

Ramos, J., Sendra, J., Sánchez, A., & Mena, A. (2017). Influence of core affect in the differential efficacy of a personality disorder intervention program. *The Spanish Journal of Psychology, 20*, 5. doi:10.1017/sjp.2017.2

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805.

Russell, J., Weiss, A., & Mendelsohn, G. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502. doi:10.1037/0022-3514.57.3.493

Rusting, C., & Larsen, R. (1997). Extraversion, neuroticism, and susceptibility to positive and negative affect: A test of two theoretical models. *Personality and Individual Differences*, *22*(5), 607-612. doi:10.1016/S0191-8869(96)00246-2

Swinscow, T. D. V., & Campbell, M. J. (2002). *Statistics at square one* (pp. 111-25). London: Bmj.

TIIM (The Incredible Intervention Machine) – BMS LAB Hight-Tech meets Social Sciences. (2018, September 26). Retrieved May 4, 2019, from https://bmslab.utwente.nl/knowledgebase/tiim/

Van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, *50*(6), 93.

Versluis, A., Verkuil, B., Lane, R. D., Hagemann, D., Thayer, J. F., & Brosschot, J. F. (2018). Ecological momentary assessment of emotional awareness: Preliminary evaluation of psychometric properties. *Current Psychology*, 1-9.

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.

Zirkel, S., Garcia, J. A., & Murphy, M. C. (2015). Experience-sampling research methods and their potential for education research. *Educational Researcher*, *44*(1), 7-16.

Appendix A

12 neuroticism items of the EPQ-S:

1 Does your mood often go up and down?

2 Do you ever feel ‘just miserable’ for no reason?

3 Are you an irritable person?

4 Are your feelings easily hurt?

5 Do you often feel ‘fed-up’?

6 Would you call yourself a nervous person?

7 Are you a worrier?

8 Would you call yourself tense or ‘highly strung’?

9 Do you worry too long after an embarrassing experience?

10 Do you suffer from ‘nerves’?

11 Do you often feel lonely?

12 Are you often troubled about feelings of guilt?

Appendix B

E-mail “Start of the Study”:

Thank you for taking part in our study.

In the following, we will provide a short outline on the set-up and the theoretical background of this study. In general, this study serves the investigation of changes in core affect over time.

Core affect represents what is commonly called mood or feelings, and consists of two dimensions- valence and arousal. **Valence** represents how **pleasant** you feel at the moment and ranges from **unpleasant** to **pleasant**. **Arousal** represents how **activated** you feel at the moment and ranges from **sleepy** to **activated**. These two dimensions are represented in the **Affect Grid**, the test you will fill in repeatedly in the course of this study. The Affect Grid is a coordinate system with *valence represented on the x-axis* (horizontal) and *arousal on the y-axis* (vertical). You can refer to the point where both axes cross as *'neutral state of feelings'*. By setting a mark somewhere in this coordinate system you indicate how you feel- both pleasant/unpleasant and sleepy/activated in one mark! (this means, that if you feel very active and pleasant you set a mark far to the right upper corner if you feel pleasant but rather sleepy/inactive you set it far to the right but in the lower corner)

(please read the information above carefully- understanding this is necessary to successfully fill in the affect grid! you can also make a screenshot of this description in case you want to read it at a later point again)

Your task is *to set a mark in the coordinate system every two hours between 10:00 and 20:00 o'clock* (six times per day) for 7 days. You do not have to keep track of the time yourself, the app will send you a notification every two hours within this time frame to fill in the Affect Grid. *(therefore please allow the app to send you notifications!)*

If you are not able to fill in the grid right away, you can do so a few minutes later, but please try to do so in time. We hope that you can answer as many times as possible!

To start the study, please download the app TiiM - The incredible intervention machine and log in with the credentials you chose. You will start by answering three questionnaires and then the study will begin-

We hope you are as excited about this study as we are!

Best,

Julia Müller, Laura Stevens, and Tim Tiede.

PS: In case of any open questions, issues, or critique please do not hesitate to contact

Julia Müller (j.muller-1@student.utwente.nl)

Laura Stevens (l.s.stevens@student.utwente.nl)

or

Tim Tiede (t.tiede@student.utwente.nl)

Appendix C

Descriptive Statistics on Arousal and Valence per Individual

ID		N	Minimum	Maximum	Mean	Std. Deviation
1	Valence	46	-64	92	42,78	44,239
	Arousal	46	-86	89	9,85	54,646
	Valid N (listwise)	46				
2	Valence	42	-85	77	,40	37,911
	Arousal	42	-85	83	9,69	40,344
	Valid N (listwise)	42				
3	Valence	46	-15	99	52,37	31,726
	Arousal	46	-100	71	,61	44,929
	Valid N (listwise)	46				
4	Valence	48	-80	85	-5,13	38,038
	Arousal	48	-84	66	-10,96	33,102
	Valid N (listwise)	48				
5	Valence	48	-19	52	22,96	15,319
	Arousal	48	1	47	25,31	10,362
	Valid N (listwise)	48				
6	Valence	42	-100	83	12,57	52,526
	Arousal	42	-100	92	,10	51,083
	Valid N (listwise)	42				
7	Valence	42	-51	74	21,48	26,901
	Arousal	42	-55	63	4,57	30,745
	Valid N (listwise)	42				
8	Valence	47	64	100	91,77	8,981
	Arousal	47	-60	100	78,00	37,514
	Valid N (listwise)	47				
9	Valence	46	1	44	21,17	10,732
	Arousal	46	-64	63	9,07	23,968
	Valid N (listwise)	46				
10	Valence	28	-23	86	34,54	30,498
	Arousal	28	-100	59	5,25	32,097
	Valid N (listwise)	28				
11	Valence	29	-24	97	50,79	30,137
	Arousal	29	-77	89	22,97	48,045
	Valid N (listwise)	29				
12	Valence	42	-51	74	11,33	31,223
	Arousal	42	-100	79	14,60	39,941
	Valid N (listwise)	42				

13	Valence	38	-93	88	45,03	42,914
	Arousal	38	-84	77	6,08	47,255
	Valid N (listwise)	38				
14	Valence	42	-83	89	36,64	32,062
	Arousal	42	-89	73	3,79	44,544
	Valid N (listwise)	42				
15	Valence	42	-53	66	22,71	27,521
	Arousal	42	-67	76	22,79	37,425
	Valid N (listwise)	42				
16	Valence	42	-54	91	54,10	26,753
	Arousal	42	-47	76	23,02	29,581
	Valid N (listwise)	42				
17	Valence	42	-46	73	33,76	27,300
	Arousal	42	-79	91	26,93	42,147
	Valid N (listwise)	42				
18	Valence	42	-66	64	12,64	32,773
	Arousal	42	-46	62	13,02	26,204
	Valid N (listwise)	42				
19	Valence	41	1	100	77,46	32,300
	Arousal	41	-86	75	3,98	42,476
	Valid N (listwise)	41				

Appendix D

Descriptive Statistics on Neuroticism Scores by Individual

ID		N	Minimum	Maximum	Mean	Std. Deviation
1	Neuroticism	12	1	5	3,58	1,379
	Valid N (listwise)	12				
2	Neuroticism	12	1	4	2,50	1,000
	Valid N (listwise)	12				
3	Neuroticism	12	1	5	2,75	1,055
	Valid N (listwise)	12				
4	Neuroticism	12	2	5	3,67	,888
	Valid N (listwise)	12				
5	Neuroticism	12	2	5	2,92	,900
	Valid N (listwise)	12				
6	Neuroticism	12	1	5	3,33	1,435
	Valid N (listwise)	12				
7	Neuroticism	12	2	4	3,17	,835
	Valid N (listwise)	12				
8	Neuroticism	12	1	3	1,33	,651
	Valid N (listwise)	12				
9	Neuroticism	12	1	3	1,83	,577
	Valid N (listwise)	12				
10	Neuroticism	12	1	3	2,25	,754
	Valid N (listwise)	12				
11	Neuroticism	12	2	4	2,75	,622
	Valid N (listwise)	12				
12	Neuroticism	12	2	5	2,83	,937
	Valid N (listwise)	12				
13	Neuroticism	12	1	5	3,08	1,084
	Valid N (listwise)	12				
14	Neuroticism	12	1	5	3,00	1,128
	Valid N (listwise)	12				
15	Neuroticism	12	1	2	1,33	,492
	Valid N (listwise)	12				