Mobile Mood Tracking: An Investigation of Concise and Adaptive Measurement Instruments

HELMA TORKAMAAN, University of Duisburg-Essen
JÜRGEN ZIEGLER, University of Duisburg-Essen

Commonly used mood measures are either lengthy or too complicated for repeated use. Mood tracking research is, therefore, associated with challenges such as user dissatisfaction, fatigue, or dropouts from studies. Previous efforts to improve user experience are mostly ambiguous concerning their validity and the extent of improvement they provide (e.g., compared to established measures, such as PANAS). This paper investigates the shortening of a self-reported mood measure using smartphones with four independent samples, and provides a baseline for comparing the usability and accuracy of future measures. It first examines whether user self-assessment of overall positive and negative activations with a *two-item measure* can capture mood as well as I-PANAS-SF. It next examines user's learning effect in repeated usage of the measure. Finally, it introduces the design of an adaptive mood measure that reduces the number of questions based on its prediction of user mood fluctuations. This *adaptive measure* can potentially capture specific mood states, as well as overall mood. The paper then explores user satisfaction and compliance with this measure in a longitudinal study. The results of this paper reveal that the investigated *two-item measure* is a valid and reliable tool for capturing a user's overall mood and mood fluctuations. The negative activation from this measure is associated with stress. Our results suggest that the association between mood and stress generally depends on the measure of mood and its items. We discovered that a non-complex self-explanatory measure is fairly resilient for repeated use with respect to the required effort and the accuracy of the measure in both daily and weekly evaluations. Adaptively reducing the length of a mood measure does not seem to impact user compliance but may slightly improve usability. We also noticed that positive and negative activations have a slightly different pattern of behavior with reference to the preceding mood states.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Empirical studies in interaction design*; *Human computer interaction (HCI)*; **Usability testing**; *User studies*; *Ubiquitous and mobile computing systems and tools*; • **Information systems** → *Personalization.*

Additional Key Words and Phrases: Mood Tracking; Interactive Design; Stress; Smartphone; User Compliance; Emotion; Affect

ACM Reference Format:
Helma Torkamaan and Jürgen Ziegler. 2020. Mobile Mood Tracking: An Investigation of Concise and Adaptive Measurement Instruments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 155 (December 2020), 30 pages. https://doi.org/10.1145/3432207

1 INTRODUCTION

Mood, studied mostly using mood tracking, influences human judgment, decision making, thoughts, and perception. It plays a prominent role in health and well-being, as well as in human interactions and behavior. In 2014, mood tracking was reported to be among the top five items of interest for self-tracking in the quantified-self community [14]. Detecting, modeling, and predicting mood is crucial for building effective health-related and mood-aware systems. Mood has traditionally been estimated based on either pen-and-paper questionnaires or

Authors' addresses: Helma Torkamaan, University of Duisburg-Essen, Forsthausweg 2, 47057, Duisburg, Germany, h.torkamaan@acm.org; Jürgen Ziegler, University of Duisburg-Essen, Forsthausweg 2, 47057, Duisburg, Germany, juergen.ziegler@uni-due.de.

interviews, and tracked using Experience Sampling Method[1] (ESM) [62]. With the ubiquitous use of smartphone technologies as research tools for tracking and modeling human behavior and running ESM, e.g., see [48, 70, 79], researchers are more often using smartphones to measure mood and emotion, as well. Self-reported measures are the standard method for capturing mood. Today, researchers use smartphones to try to implicitly detect or automatically predict mood fluctuations in order to overcome limitations in traditional methods, such as user fatigue in ESM. However, they still need to rely on self-report-based mood measures to train and validate their models. Therefore, proper smartphone-based self-reported mood measures are the cornerstone for measuring, understanding, detecting, modeling, and predicting mood.

The length and quality of a smartphone-based mood measure can impact user experience, fatigue, and compliance, as well as response quality, in an ESM study. The challenges of running in situ measurements and satisfying user expectations of an app, even for an ESM study, has compelled researchers to improve the user interface (UI) of and the interaction with an app based on a classic mood measure, and to prefer shorter measures of mood in general. Accordingly, researchers sometimes modify, shorten, or redesign classic measures, or even devise new measures of mood. However, modified or newly devised measures should be validated before being used. Using invalid measures can lead to arbitrary mood data, incorrect models, inaccurate detections, and unreliable predictions. Accordingly, to what extent, if at all, shortening or altering a traditional mood measure for smartphone users influences its validity and user experience, is an intriguing issue. The present paper aims to address the gaps in prior works in the field by emphasizing balancing the validity of the assessment and user experience for mood tracking, and investigating the shortening of a classic measure. This could be a useful aid and guideline for all researchers who conduct interdisciplinary research and use questionnaire-based measures.

Background and Definitions. From the early works on human affect [20, 33] to the latest research, e.g., [19, 43, 48], mood has been studied using various underlying theories, definitions, and validated measures. Although there is no one clear definition for emotion [32] or mood, these concepts are clearly differentiated in affect literature by their properties [7, 23, 59, 77]. These properties can explain why mood research is more self-report–oriented than emotion research. *Mood*, in general, is an affective state that lasts anywhere from minutes to hours or even days. Due to its longer duration, mood seems to have an influence on, and correlation with, various other behavioral or psychological phenomena, such as cognition, memory, and stress. Mood does not emerge suddenly, and an individual experiencing a certain mood may not be able to identify its cause. As a consequence, it is harder to capture reactions or responses, such as physiological responses for mood; therefore, mood assessment relies mainly on subjective self-reported experiences. *Emotion* is a mental process different from mood. It is shorter —anywhere from a few seconds to minutes— occurs suddenly due to stimuli, and has a cause and observable signals. Unlike mood, one can measure emotion using its signals, such as specific physiological responses or facial expressions. We use the term *affect* as a broad term for feeling states, either mood or emotion. *Mood measure* in this paper refers to self-reported mood questionnaires or instruments that are used to capture or measure mood states of an individual. Below, we define two more concepts: assessment quality and app quality. The core of these concepts is not new; however, naming and formalizing them with technical terms allows us to properly discuss the limitations of existing work and the contributions of this paper.

Assessment quality refers to a mood measure's validity, reliability, and relevance of construct and theory, as well as its comparability to mood measures used in previous studies. Every mood measurement accordingly has to have an adequately high level of assessment quality. The *assessment quality* can be ensured by using relevant, valid, and reliable classic measures. For a high quality of assessment, it is essential to note within which contexts and with which limitations a measure of mood — including general-purpose or for specific mood states — has been used

[1]Experience Sampling Method (ESM) or Ecological Momentary Assessment (EMA) is repeated, mostly trigger-based (event, schedule, or random) sampling of users' state, feeling, behavior, and thoughts over time. Such samplings can be conducted as seldom as once every few weeks or as frequently as several times per day.

before relying on the measure, or its resulting associations of mood with other phenomena. *Assessment quality* can also be ensured by validating newly devised measures, e.g., against relevant classic measures, and indicates the correctness of a measurement. It also represents the extent to which a captured mood value can be relied on and is accurate, relevant, valid, and usable for investigating mood or other associated phenomena, or building behavioral or predictive models. Mood tracking with smartphones, like with any other mood measurements, should have an acceptable assessment quality. However, it requires more considerations.

When using smartphones to measure mood, it is essential to have a mood measure application (app) that is usable, user-friendly, and that meets user expectations. Users have high expectations of smartphone apps. An app used in a study should meet those expectations and have a high level of usability in order to engage participants throughout the study and to capture valid inputs. Therefore, a successful mood tracking app, and corresponding study, has to address *app quality* in addition to *assessment quality*. *App quality* referes to the user experience, usability, and usefulness of a smartphone-based mood measure, which in the long run, can potentially improve user compliance and decrease user fatigue, particularly in ESM mood tracking. Researchers can ensure adequate app quality by using human-centered design, evaluating the user experience of the app, and testing functional and structural qualities of the app as software.

The longitudinal assessment of mood or mood tracking is different from a one-time assessment. Due to the long-lasting nature of mood and the long-term nature of any human behavior tracking, mood tracking requires frequent assessments over time. In mood tracking, a user would have to interact with a mood measure repeatedly, even several times per day. This repetition could add additional challenges, such as user compliance and learning effects, to the design of a mood tracking smartphone app. Paying attention to the app quality can address these challenges to some extent. However, focusing on achieving high app quality may influence the assessment quality, particularly in a classic measure.

Motivations. To make a smartphone-based mood tracker, researchers can either use a relevant classic mood measure and transfer it into an app, or construct a new measure. Transferring a classic mood measure into an app could result in several challenges related to app quality, as well as to assessment quality. Classic measures of mood were designed and validated in the past using pen-and-paper forms (usually depicting Likert-type scale) rather than smartphones. They are either too lengthy or complex for everyday repeated use. A typical classic mood measure can have up to 132 items [81]. Considering the screen-size limitations and distracting nature of smartphones, it would be unpleasant for users and impractical for researchers to use these lengthy classic measures repeatedly without any modification or shortening. Changing the design of or shortening a measure could, however, potentially influence the quality of the assessment or validity of a measure. Therefore it is vital to assess the extent of loyalty to the original measure and to use proper validations where necessary. Researchers can also devise a measure that is designed specifically for smartphones, and thus has a high *app quality*. Such a devised mood measure would still needs to be evaluated for its *assessment quality* before it could be used reliably. However, many mood tracking apps have no such validations or assessments.

Despite the importance of both app quality and assessment quality for an accurate, valid, and useful measurement of mood, studies using mood tracking commonly address either *app quality* or *assessment quality*, or even neither. A group of these studies measures mood not as the primary variable of the study, but rather in combination with other variables, e.g., see [31, 63]. Another group focus only on specific mood states instead of a general assessment of mood, e.g., see [79]. Other studies include those investigating general assessment of mood or discussing a new measure of mood, e.g., see [27, 37, 53]. However, most studies in these groups — due probably to their specific foci and research questions — provide neither a comparison of their measures with existing mood measures nor a proof of validity or reliability for devised or modified measures. One study tried to address both app quality and assessment quality for its devised measure, the Photographic Affect Meter (PAM) [50]. This measure was built to be pleasant for users, short enough for in situ assessments, and to reliably and quickly measure emotion using a smartphone. However, despite the efforts of the designers, the construct of this measure

seems to focus more on app quality than assessment quality, and as a consequence, it does not provide a complete measure of the affective space. We explain this limitation further in the related work section (section 2).

A desirable feature for researchers when choosing a measure for mood tracking and ESM is the briefness of the measure. A drawback of accurate classic mood measures frequently mentioned in the literature [37, 44, 50], is their length. Unfortunately, today, one can hardly find a measure that is brief and adequately addresses both app quality and assessment quality. For the most part, there is a trade-off in mood tracking between assessment quality and app quality. It has not yet been established how one can improve the user experience or compliance in their studies by using classic measures. Classic transferred measures of mood have not been explored for their usability on smartphones in an ESM study either. It is therefore unclear how users perceive classic measures of mood, and no baseline currently exists with which improvements of a modified or newly devised measure can be compared. It is still unknown how the method of shortening a measure may impact its assessment quality and app quality in longitudinal assessments of mood. Altogether, considering the importance of self-reported mood measures, the significance of having tools compatible with classic measures, the challenges of transferring classic measures to smartphones, and the increasing interest in using short measures, exploring the use of classic measures and their alternatives in smartphones alongside app usability is a worthwhile endeavor.

Contributions. To address the above-mentioned challenges, we leverage information from past works that focused merely on app quality, by concentrating on assessment quality, as well. In this paper, we reduce the length of a general mood measure while balancing the app quality and assessment quality. We investigate if one can measure mood based on an important classic measure of mood, the Positive and Negative Affect Schedule (PANAS) [75], with only two overall questions, and we also consider the learning effect in the repeated use of such a measure. Our intuition is that such a shortened measure is still better than an arbitrary measure of mood or invalid measures. We find the usability and user compliance of a basic classic measure (I-PANAS-SF) when transferred to smartphones, and compare it with further modifications. In particular, we learn whether a measure adaptive of a length, that can possibly learn from user behavior in the future, could be a good solution for ESM. A key question, then is: how well would such a system improve user experience of and compliance with a measure? In order to achieve these goals, we developed an app for the study that implements a variety of mood measures[2]. We used the app to investigate various dimensions that influence the usability and user compliance of a mood measure, such as repetition and the number of questions.

In addition to mood, we also take a look at the association between mood and self-reported stress. As mentioned earlier, mood is a phenomenon that is related to other phenomena, such as stress. The study of stress is usually associated with a measure of affect, mood, or emotion. Lazarus [40] even says that the three concepts of emotion, coping, and stress should be studied togather as part of a single conceptual unit. ESM-based studies in the field often track mood and stress together, e.g., [35, 43, 65, 67]. Tracking stress, in addition to mood, allows researchers to consider the influence of these phenomena on one another and help building predictive systems in the future. In this paper, our focus is on the mood measures; however, by capturing the self-reported experience of stress, we can further evaluate our mood measures against their classic versions in the literature, and look into its strength in reflecting the association of mood and stress, as well.

In summary, this paper, by focusing on reducing the length of the general mood measure, supports the community with the following contributions:

(1) Encouraging the community to consider both app quality and assessment quality in longitudinal studies, and highlighting issues that could arise when improving usability without proper validation
(2) Presenting a baseline from classic measures for comparing and building future measures in the community and providing its source code and dataset

[2]The application and its source code are open for non-profit academic research purposes and can be requested here https://torkamaan.de/#download or https://interactivesystems.info/developments/pax-mood-tracker

(3) Balancing app quality and assessment quality by investigating the shortening of a classic measure and presenting an example short instrument for mood tracking in the field, which performs well in user experience and outperforms existing measures in quality of assessment
(4) Showing the validity and reliability of the instrument with various evaluations
(5) Introducing and discussing an adaptive ESM approach and finding out how adaptively shortened number of questions may improve usability and user compliance

The rest of this paper is organized as follows: Section 2, gives an overview of the related work, followed by a list of the research questions. Section 3 provides a description of the app used as a tool for conducting the studies, and describes the study inclusion criteria and procedure. Section 4 explains the datasets and sample description of this study, which are used to address all the research questions. Sections 5, 6, and 7 each focus on one of our three main research questions. Some of these sections outline more details on the research questions-specific method, and then present the results and discussions. Finally, section 8 discusses further implications and limitations of our findings and section 9 summarizes our conclusions.

2 RELATED WORK AND RESEARCH QUESTIONS

As we briefly mentioned in the previous section, section 1, the majority of the works related to mood tracking in the field, primarily address *app quality*. In general, studies that use mood tracking can be categorized into three groups. The first group include studies, e.g., [4, 13, 43, 65, 67], that use a self-defined arbitrary measure of mood, and capture items, such as good, bad, neutral, fair, fine, sad, and happy. The values of mood in such studies, unfortunately, cannot be easily mapped to any classic measure of mood, and as a consequence, assessment quality is diminished.

The second group include studies, e.g., [16, 66, 79] that focus on discrete affective states, such as depression and anxiety and therefore, do not capture general mood. These studies may be related to specific target groups, such as mental health patients. Despite the importance of mental health-related studies, their mood-tracking scope is limited to the target group or specific disorders. Specific mood measures — different from other general mood measures — have been designed and should be used to detect and monitor symptoms of related disorders, e.g., Beck Depression Inventory (BDI) [6] and Center for Epidemiologic Studies Depression Scale (CES-D) [52]. Studies in this group might also be related to discrete models of affect in which distinct affective states are considered using classic measures, such as Profile of Mood States (POMS) [46] and Multiple Affect Adjective Check List (MAACL) [81]. Discrete models of affect, however, are essentially limited to specific affective states on which the measure concentrates and are unable to explain the interrelations of affective states or the general feelings of an individual. For instance, POMS only looks into six states of tension, depression, anger, vigor, fatigue, and confusion. Due to the limitations of discrete models to draw a complete picture of an individual's feelings and to explain high correlations between distinct affective states with similar valence — e.g., anger and sadness —, researchers have instead been using dimensional models of affect [77].

The final group include studies, such as [37, 44, 48, 50, 61, 63, 72], that rely mainly on dimensional models of affect and measure general mood with relevant dimensions. Dimensional models of affect usually define the affective space with a two-dimensional space (sometimes three-dimensional or even more depending on the theory), where every affective state can be explained by its relative position in this space. Two leading classes of these models are the dimensions of pleasantness-energy[3] [55, 56] and Positive Affect (PA)-Negative Affect (NA)[4] [76]. Three commonly known classic measures of pleasantness-energy dimensions are Mehrabian and Russell [47]'s scale, Affect Grid [57], and Self-Assessment Manikin (SAM) [9]. PA and NA are mainly measured using PANAS [75], or its extension, PANAS-X [74], or one of its shorter forms, namely International-PANAS-Short

[3]Sometimes also called pleasure-arousal or pleasure-engagement
[4]Sometimes also called positive activation-negative activation

Form (I-PANAS-SF) [69]. It has been argued that PA-NA dimensions are only a 45° rotation of pleasantness-energy dimensions [57, 75]. However, despite this theoretical explanation for orientation of these dimensions, PA-NA scores are *not* fully mappable to pleasantness-energy scores [29, 57], particularly for scores of energy, in which the correlation of NA and energy is negligible.

The scientific community have frequently used either PA-NA or pleasantness-energy dimensions to investigate affect, mood, and emotion. But, the amount of evidence and support for these dimensions is not the same for every affect concept. For example, pleasantness-energy dimensions are often discussed to explain emotion, affect, or an individual's attitude toward an object, stimuli, or image both in the affect literature [9, 57], and the affective computing community [12, 18, 51]. There seems to be more evidence for the evaluation and applicability of PANAS based on PA-NA dimensions, in capturing mood than measures, such as Affect Grid, that are based on pleasantness-energy dimensions [77].

Most of the dimensional measures are either too lengthy or complex to learn. Although complex but short non-self-explanatory measures can be used more quickly after a training phase, one cannot know for sure without the training if the captured self-reports correctly reflect user mood states. As a consequence, it would be an additional burden to a research study to ensure that the training phase has been completed correctly, or even, conducted in person. Moreover, the majority of dropouts from a study occurs within the first usage instances. On this account, the recorded values from a complex measure at the beginning of a study cannot be relied upon, since users may not know how they should answer the questions. Affect Grid is a widely used example of a short but non-self-explanatory measure that has been used in the community, e.g., in [61]. This measure has a very long instruction text with at least six examples, each with descriptions [57]; in comparison, PANAS has only one or two lines of introductory text. After evaluating the assessment quality of Affect Grid, Killgore [38] argued that it is only a *moderately* valid and reliable measure.

Non-complex and self-explanatory measures with short instructions, such as PANAS, are usually too lengthy and burdensome to be used repeatedly in a longitudinal study. PANAS is a widely used and discussed measure of mood[5]. It is a valid and reliable measure of general mood and is mainly used for non-clinical populations [54]. PANAS has 20 items (ten items each for the two independent dimensions of PA and NA), which is too long to be administered several times per day. In PA-NA model of affect, affective states that convey excitement and pleasure, such as active and excited, are considered as high PA, and states like sleepiness fall on the low PA. Similarly, affective states like calm and relaxed are considered as low NA, whereas hostile or nervous are high NA. PANAS only measures items of high PA and high NA, and it is validated for mood tracking even with large populations [17]. Mood extracted using PANAS has been studied frequently in the literature and found to have correlations with other psychological phenomena. For example, Watson [73] found a correlation of 0.44 between NA and self-reported single-item stress. Similarly, Crawford and Henry [17] found a correlation of .67 between NA and stress measured via depression, anxiety, and the stress scale of Lovibond and Lovibond [45]. NA is correlated with depression and anxiety [17], whereas PA is associated with social interaction [75].

Due to the vast body of support, validation, and application of PANAS in the literature, there has been an increasing interest in developing new valid and reliable measures based on PANAS that are shorter, have fewer items, and can be applied in cross-cultural or international populations. For example, Kercher [36] worked with a shorter version of PANAS, reducing the number of items to 10: five each for PA and NA. Thompson [69] argued that the shorter version of Kercher [36] contains inter-correlated items and can be further improved. He developed yet another shorter version of PANAS with 10 items, calling it I-PANAS-SF, and evaluated it with various international populations [69]. I-PANAS-SF was later used and evaluated in other studies [34]. PANAS and I-PANAS-SF are both valid and reliable measures that can give an accurate estimation of the user's overall

[5]It is one of the most (if not the most) discussed and widely used measures with more than 37, 000 citations according to Google Scholar [1], 17, 000 on Web of Science [3], and 4, 000 on PubMed [2].

mood and are proper measures for mood tracking. Nevertheless, despite the efforts to reduce the number of items, the ten-item questionnaire is still too long for frequent use in an app, and the number of questions a user must answer throughout a study can proliferate to an unmanageable amount. This challenge has been a motivation for the community to investigate newly devised measures instead [50].

In an effort to address both the dimension of *app quality* and study requirements, some researchers devised their own versions of mood measures. These devised mood measures need to have a high *assessment quality* in order to be trusted by other researchers. An example of such an effort is PAM [50], which is a rare example of a devised measure in the field that tries to balance app quality and assessment quality. This measure, however, has a few shortcomings. A serious limitation is related to the construction of this measure, and the mapping of pleasantness-energy dimensions to PA-NA dimensions without considering the difference between these dimensional models. PAM orders a set of images based on the level of pleasantness and energy values, and then maps their positions to a grid similar to Affect Grid, with a total score between 1 to 16 and a value between 1–4 (or -2–2) for pleasantness and energy dimensions. However, despite having a design based on pleasantness-energy dimensions, this measure chose to validate with PANAS instead of using SAM or Affect Grid. This choice does not seem to have any proper theory-based reasoning, besides assuming that PANAS is a more popular measure and a summary of other measures. PAM score has a significant correlation of .71 with PA and a weak correlation with NA [50], despite PAM having negatively loaded items (pictures). Both dimensions of affective space are needed to fully describe affective states and capture general mood. PAM consequently cannot adequately capture NA. Furthermore, PAM was designed and tested with the instruction to capture feelings in the moment (i.e. *"how you feel right now"*), and therefore, using this instruction, it is unclear whether PAM measures emotion or mood. Altogether, PAM does not seem to be a measure with an adequate assessment quality for tracking general mood over time.

Another example of a devised mood measure is the pictorial design of Desmet et al. [21]. It captures eight distinct affective states, containing low PA and NA items. Although user-friendly, this design has not been validated nor compared with classic measures. Other measures, e.g., [24, 27, 53], similarly, do not provide sufficient evidence of assessment quality while focusing on app quality. Khue et al. [37] also concentrated only on app quality and clearly indicated that assessment quality was outside the scope of their paper. As a consequence, the captured mood values and, accordingly, the developed models based on any such measure cannot be compared to, or reproduced across, various studies. The limitations of a measure to provide a proper assessment quality brings into question the effectiveness of research studies based on that measure for tracking mood, particularly in relation to neglected dimensions of mood.

Despite efforts to design mood measures for smartphones, current solutions are inadequate in providing a measure comparable to PANAS. Alternative dimensional models, such as pleasantness-energy dimensions, have been argued to reflect on the same concept of mood, yet they cannot be used as a replacement for PA-NA dimensions [29]. If researchers need to stay loyal to PANAS, and correspondingly, its underlying theoretical model of affect, it would be unclear to what extent they can modify or transfer it to a smartphone app suitable for sustained use and still have a valid assessment. The user-friendliness and compliance with this measure over time are also unknown, despite its potential to be used as a baseline. This paper specifically answers the following research questions:

- **RQ1 – Two-item overall questions**
  - (i) Can user self-assessment of overall feeling reflect mood? Specifically, would a short two-item questionnaire about overall feeling give a determination of overall PA and NA captured via PANAS and represent the mood or its fluctuations?
  - (ii) Is such determination of mood with the two-item measure different in day-to-day measurement, versus one-time or weekly assessment of the mood?

(iii) Considering the importance of NA in the study of stress, is such an overall determination of NA associated with a user's self-reported stress?

- **RQ2 – Learning effect**
  (i) Does repeated use of the mood measure teach users to better and with less effort reflect on and report their mood using fewer questions?
  (ii) Is repeated use of the mood measure, and its possible effect on user familiarity with the measures, different in day-to-day versus weekly repetition?
- **RQ3 – Adaptive design**
  (i) How can we design a mood measure that gives a more complete picture of the user's mood by capturing both the overall mood and a certain level of detail on the mood states, while keeping user effort minimal?
  (ii) Do fewer questions (adaptively reduced) affect the *(a)* usability or *(b)* compliance?

3 STUDY PROCEDURE AND PLATFORM DESCRIPTION

To investigate the research questions, we developed an Android application (app) as a platform for experience sampling with the following capabilities: utilizing a variety of designed mood trackers; assigning users randomly to the study conditions; and storing the responses on our server. These capabilities allowed us to explore the research questions with different samples. The app has the following components: pre-study questionnaire, twice-daily sampling, weekly sampling, post-study survey, user feedback, and settings (figure 1).

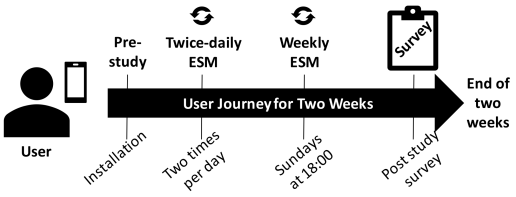

Fig. 1. User journey of participants

3.1 Pre-study Questionnaire

Following the installation, the app starts with an onboarding sequence that users cannot skip, in which they are asked basic demographic questions. Within this sequence, we present a description of the study and inform users (participants) of their data protection rights, according to GDPR. The app then shows users the pre-study phase which has a few optional tasks or questionnaires. The pre-study aimed at within-subjects comparison of several mood measures (some outside the scope of this paper), and users record their mood with all measures with a random order. The pre-study can only be filled out once.

3.2 Experience Sampling

Our experience sampling app is designed for both pre-scheduled, pop-up based mood entries, and on-demand mood entries. For pre-scheduled entries, the app pushes the respective mood questionnaire to the user's phone screen together with an ongoing notification at least two times per day, as long as the app is installed on the phone. If a user does not engage with the mood sampling pop-up page (either by dismissing, completing, or exiting in the middle), the ongoing notification will remain active until the time-frame for the entry passes. The first compulsory mood sampling is activated between 9:00 and 12:59 (default: 09:00), and the second entry between 14:00-15:59 (default: 14:00). The participants can add a third voluntary scheduled sampling event between

19:00 and 21:59 using the app settings. We chose these time frames to capture mood fluctuations throughout the day. The mean of a PA value is generally lowest in the morning, and then seems on average to rise and reach a maximum before 15:00, staying stable until 21:00 [15, 49]. Therefore, these time frames would be likely to capture values close to minimum and maximum PA in a day. NA values, according to Clark et al. [15] appear to stay stable throughout the day. Using the settings, users can modify the timing of the two compulsory samplings within the given time-frame limits. Users can also enter and record their mood manually (on demand) using a button on the main screen of the app. In addition, the app has a weekly mood assessment scheduled for Sundays at 18:00 that specifically asks for the mood *during the past few days*. Each sampling event is recorded in the database and tagged as either *completed*, *dismissed* (i.e., if a user does not engage with sampling and dismisses the sampling event), or *invalid* (i.e., when the user starts answering but does not finish the sampling event). In addition to the mood values, the app records the timestamps of the sampling event, the timestamp of when users start responding, and the submission timestamp.

3.3 Mood Measures

We designed and implemented various mood measures from different theoretical models, of which only three mood measures are used to answer our research questions[6]. Users were randomly assigned to one of these measures or a combination of them, and used the assigned measure(s) on a twice-daily basis throughout the entire study. They also received a weekly overall mood sampling; however, the chosen measure for weekly ESM, which we examine in this paper, was the same for all users, regardless of their assigned group. Each mood measure is equipped with a help button and guidelines according to the classic measures' original guidelines, which instruct users on how to answer the questions to record their mood.

This paper deals with a total of three mood measures and a self-reported stress measure. The first measure (Q-10) is a 10-item questionnaire and represents the classic measure I-PANAS-SF [69]. We presented the Q-10 items in two tabs (figure 2.a): randomly ordered items of PA; and randomly ordered items of NA. PA items are Alert, Inspired, Determined, Attentive, and Active; and NA items are Upset, Hostile, Nervous, Afraid, and Ashamed. In the original I-PANAS-SF [69], scores of PA and NA items are aggregated separately to get the overall score of PA and NA. In Q-10, we averaged the items' scores, finding final PA and NA values between 1 and 5.

The second measure (Q-2) asks users to enter their self-rated overall positive and negative activation values (figure 2.b). Q-2 includes a help button that shows additional information on request. Users could choose a value between labels to have more flexibility when needed. This may have helped obtain values that better reflect the calculated PA and NA scores from Q-10 (continuous values between 1 and 5). The question used for all measures was the same and, is depicted in figure 2. For the weekly assessments, however, the question was slightly different, and the phrase *for the past few hours* was replaced by *for the past few days*.

The third measure has an adaptive design. The goal of mood trackers is primarily to detect and model mood fluctuations. While a valid, short measure like Q-2 might return a determination of the fluctuations, it cannot capture the details of a user's state as well as a full measure like Q-10. For example, if we know from Q-2 that user NA has increased, we won't know without Q-10 whether it was due to items like afraid, hostile, or ashamed (mood states). Using an adaptive design, that customizes or selects the questions, both mood fluctuation and user mood states could be captured using fewer questions and with less effort, while keeping the interaction pleasant and not necessarily long or boring. However, before implementing advanced algorithms, we wanted

[6]Other measures were part of a bigger project intended for different research questions that are outside the scope of this paper. These research questions concentrate on investigating different theoretical models of affect, such as hierarchical model of affect [68] and exploring various user interactions and interface designs. Overall, 1, 300 users in the whole project installed the app and were randomly (between-subjects design) assigned to nine study groups and conditions (some unrelated to this paper). One-third of our users kept using the app after two weeks (up to more than 245 days ($M = 27.847$, $SD = 44.162$, $Mdn = 8.46$) until October 2019). The related materials and publications can be accessed under authors' websites upon publications.

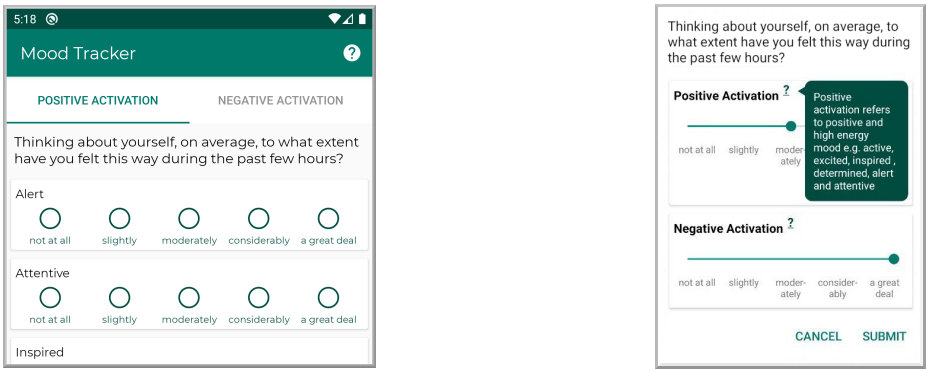

Fig. 2. **(a)** On the left: I-PANAS-SF (Q-10) was presented in a page with two separate tabs of positive activation and negative activation. The screenshot is cropped. **(b)** On the right: Two-item overall assessment of positive and negative activation (Q-2) was presented in a dialogue.

to first investigate whether an adaptively shortened measure would lead to any improvement, such as a high assessment quality, a high app quality, or in particular, improved user satisfaction. Therefore, we came up with a basic algorithm as the adaptive design. We describe this algorithm in more detail in section 7, and evaluate it to answer RQ3.

The final measure used in this paper is a self-reported stress measure, similar to [73]. At the end of each mood sampling event, we asked for the user's perceived feeling of stress in a separate dialogue. This question targets users' experience of stress and asks users to specify the extent to which they felt stress with a five-point Likert-type scale. Daily hassles and demands can cause stress, which plays an important role in natural mood fluctuations during the day. With this measure, we can investigate and compare the association between mood and stress between our three mood measures.

3.4 Feedback, Navigation, and Post-study Survey

The main screen of the app shows a mood report (feedback) to users. The report has two options: the current day and all days of the week. The mood state of each day is shown as a smiley face (figure 3) that can have five different values, depending on the overall mood. The overall mood of a day is calculated by averaging the recorded mood values of that day. In the case of no mood entry for a day, the face appears as an empty, gray circle. Using the menu, users can observe a report on their entire mood fluctuation history (captured from daily entries) – a diagram showing their mood entries for the whole period of the app life. The same navigation menu has an option for the post-study survey. A scheduled event opens the survey page on the user's phone two weeks after the installation. The survey page contains a link — combined with a unique, anonymous user identification code — to the online Soscisurvey [42] survey. The app is offered in both English and German languages and syncs all new user data four times per day to the server.

3.5 Participant Recruitment and Inclusion Criteria

After releasing the application in the Google Play Store as a mood tracker tool, we advertised the study through media promotion. Participants could access the link to the app, together with the study description on our website. The study period was announced to be for two weeks (though users could continue using the app after this period). We filtered the installation instances from the server based on our study participation conditions, which were a valid minimum age of 18 years old, an application installation of at least seven days, and at least three completed mood entries. This resulted in completed entries from 391 (out of 547) users, who were included

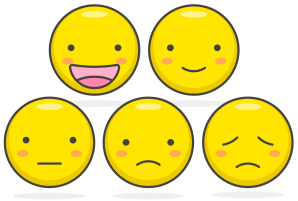

Fig. 3. Five Emojis depict users' overall mood states that were used to give a report and feedback to users on the main screen of the app. The value of PA versus NA determines the chosen emoji for a day. Twice-daily ESM users received this feedback. Emojis are taken from [41].

in the present paper. We excluded daily mood entries that were submitted after the 15th consecutive day of participation (after installation date) in order to keep a homogeneous period of participation between users. Every user was randomly assigned an anonymized user ID, data was transferred via RESTful API over SSL, and was accessible only by a local access to the server. The study was confirmed by the ethics review committee of the authors' institution before it started.

4 DESCRIPTION OF DATA SAMPLES

We used the app with various study conditions based on the described measures in section 3.3 and collected data from the pre-study, twice-daily, and weekly ESM on our server. Figure 4 depicts these study conditions. We then processed the collected data to obtain data samples for each research question investigated in this paper. To evaluate RQ1 and RQ2, we derived the following independent data samples from the pre-study and ESM data: (*A*) a diurnal (day-to-day) assessment with repeated measures of mood (at least two times per day) for each individual for two weeks; (*B*) a weekly sample assessing general mood for the past few days; and (*C*) a one-time assessment with counterbalancing of Q-2 and Q-10 across subjects. To answer RQ3, we used sample D, a diurnal mood assessment using the adaptive design, in addition to and independent of sample A. We had a post-study survey related to RQ3, in which users of samples A and D took part, but we discuss it later in section 7.1.3. Table 1 shows an overview of various samples in this paper, and below, we explain in more detail how each sample is obtained.

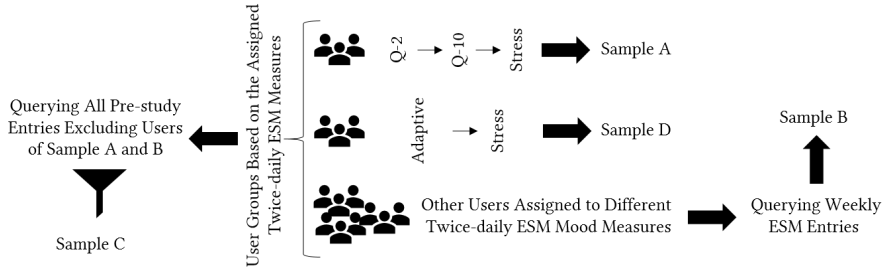

Fig. 4. Between-subject study conditions for twice-daily ESM separate users into three groups: Users who use Q-2 followed by Q-10; users who use the adaptive measure; and users who use other types of measures (not PANAS-related and outside the scope of this paper).

4.1 Sample A: Diurnal Assessment Sample

Mood is subjective, and for assessing the accuracy of a measure, one should ideally compare it with a classic measure concurrently. To assess the validity of Q-2 and its learning effect for ESM, users were assigned to a the study condition, where they first saw the Q-2, followed by the Q-10 and stress measures, for each sampling event in the app. Sample A is the resulting mood entries from the twice-daily ESM. In this sample, 90 users were assigned to the study condition, of which 37 (female: 24, male: 13, other options: 0, with a mean age of 34.27 (SD = 11.37) years) remained after applying the study inclusion criteria (described in section 3.5). Altogether, we collected 1,462 sampling events with 767 completed entries (invalid= 43, dismissed= 655). We then compared the values of PA and NA captured by Q-2 with those from Q-10, and took a look at their association with stress.

4.2 Sample B: Weekly Sample

The same mood measures — Q-2, followed by Q-10 and stress measures — were used for the weekly assessment of mood; however, the questions focused on the past few days. Of 251 users, 2,054 weekly mood entries were recorded on our server because some users continued using the app after two weeks. These entries include completed, invalid, and dismissed weekly mood sampling events. In total, we obtained 836 completed weekly entries from 182 users (female: 123, male: 56, other options: 3, with a mean age of 33.65 (SD = 11.66) years). Among these entries, 129 users completed the weekly measures for the first and second weeks consecutively. This resulted in 258 completed entries, which we considered as sample B[7]. Users of sample B interacted with Q-2 and Q-10 measures on a weekly basis. They additionally experienced a twice-daily ESM with other measures that are independent of any PANAS-related (Q-2, Q-10, adaptive, or any other measure based on them) measures or items, which therefore are not included in this study.

4.3 Sample C: One-time Assessment Sample

Every user of the app could participate in the pre-study and answer a one-time optional assessment of their mood with various measures. Taken all completed entries of the pre-study, we excluded users of samples A and B. This resulted in a total of 177 users (female: 127, male: 46, other options: 4, with a mean age of 34.66 (SD = 12.23) years). As mentioned earlier, for every mood measurement, Q-2 appears before Q-10, and thus, even in the pre-study, their order is not counterbalanced. To check if the order of Q-2 followed by Q-10 impacts our study as a limitation, we compared them with an additional measure from the pre-study, i.e. a chatbot version of Q-10. This measure was originally part of a bigger project alongside other measures. The order of the measures was counterbalanced across subjects when considering using Q-2 followed by Q-10, and a chat-based version of Q-10. Among these users, 88 were assigned to answer Q-2 before Q-10, and 89 were assigned to answer Q-2 after chat-based Q-10. In general, chat-based Q-10, although slightly different in the interface design, has a very strong significant correlation with the classic Q-10 (PA: $r(175) = .89$ ($p < .01$); NA: $r(175) = .92$ ($p < .01$)) and, therefore, we argue that it can be used reliably for this evaluation. We accordingly compared users who answer Q-2 after Q-10 to those who answer Q-2 before Q-10.

4.4 Sample D: Adaptive Design Sample

From 92 users who were assigned to the study condition of using the adaptive design at least two times per day, 48 users (female: 33, male: 14, other options: 1, with a mean age of 32.83 (SD = 11.74) years) remained after cleaning the data based on the inclusion criteria of the study, as defined in section 3.5. We call this sample the adaptive sample, or *sample D*, and it has 1,905 mood entries, among which 1,080 entries are complete (invalid= 86, dismissed= 739).

[7] 63 users (out of 251) with weekly mood entries, who were not included in sample B because of not having consecutive completed weekly entries, had previously completed the prestudy and therefore, were included in sample C.

Table 1. An overview of the described samples (A: section 4.1; B: section 4.2; C: section 4.3; D: section 4.4) in this paper.

	Sample A	Sample B	Sample C	Sample D
Number of users	37	129	177	48
Number of females	24	84	127	33
Mean Age	34.27 (SD= 11.37)	35.53 (SD=12.95)	34.66 (SD=12.23)	32.83 (SD=11.74)
Total Number of entries	1462	258	177	1905
Total number of completed entries	767	only completed included	only completed included	1080
Total number of dismissed entries	655	only completed included	only completed included	739
Total number of invalid entries	43	only completed included	only completed included	86
Used to answer	RQ1-i, RQ2, RQ3	RQ1-ii, RQ2	RQ1-iii	RQ3
Stage of the study	twice-daily ESM	weekly ESM	pre-study	twice-daily ESM

5 RQ1: TWO-ITEM OVERALL QUESTIONS — RESULTS AND DISCUSSION

To investigate the assessment quality of Q-2 and its feasibility for capturing overall mood, we examined the values of PA and NA resulting from Q-2 with those from Q-10 (RQ1-i). We inspect these values, which are recorded within-subjects in three independent samples with varying assessment types (i.e. day-to-day (sample A), weekly (sample B), and one-time (sample C)), to ensure the assessment quality of Q-2 and to explore its possible limitations (RQ1-ii). We also compared the association of mood and stress between these two measures (RQ1-iii).

5.1 Sample A: Diurnal Assessment Sample

User self-assessment of overall PA and NA seems to reflect the calculated values of PA and NA. Table 2 illustrates the description of the resulting PA and NA values of both Q-10 and Q-2 for all entries of sample A. Both PA and NA values have strong significant correlations (PA: $r(765) = .723$, $p < .001$; NA: $r(765) = .734$, $p < .001$) in Q-2 and Q-10. The PA and NA items of Q-2, respectively, show excellent internal reliability with PA and NA items of Q-10 (*Cronbach's* $\alpha = .928$ for PA and .888 for NA). There is also significant intercorrelation between NA and PA scores in both Q-2 and Q-10. The intercorrelation of PA and NA scores has been a matter of discussion in the scientific community, and researchers have reported various values from $-.23$ to $-.58$ [17, 26, 60, 68, 75]. Thompson [69] reported this intercorrelation to be higher (i.e., $-.32$) in I-PANAS-SF, as compared to PANAS. In this sample, PA and NA in Q-10 show a similar correlation to the values reported by Thompson [69], using the I-PANAS-SF scale. However, this correlation in Q-2 is significantly higher[8], which may be due to having the two Q-2 items on one page, among other reasons.

Table 2. The mean, standard deviation, and correlations of the PA and NA values across Q-2 and Q-10 in sample A.

	M	SD	Mdn	PA-Q-2	PA-Q-10	NA-Q-2	NA-Q-10	Stress
PA-Q-2	2.773	.918	2.91	1				
PA-Q-10	2.649	.946	2.60	**.723****	1			
NA-Q-2	2.221	1.024	1.99	-.546**	-.408**	1		
NA-Q-10	1.741	.830	1.40	-.447**	-.331**	**.734****	1	
Stress	2.13	1.193	2.03	-.390**	-.283**	.584**	.540**	1

The frequency and distribution of values are visualized for both PA and NA in figure 5.a. When considering this visualization and the mean values, we see that, for the majority of entries, PA and NA values of both Q-2 and

[8]Using Cocor for two non-overlapping correlations based on dependent groups [22], significant for all tests, e.g., using [64], $z = 7.187$, $p < .001$.

Q-10 are relatively close to the identity line (the red line in the graph). One can infer from the red-colored circles below the identity line in figure 5.a that the users are more likely to overrate their NA in Q-2. However, PA values – the green-colored circles – are, on average, both below and above the identity line. According to their mean values (table 2), overall, the PA captured by Q-2 is only slightly higher than that captured by Q-10. Depending on the specific values of PA and NA, Q-2 and Q-10 seem to show slightly different behavior. For instance, the value of 1 from Q-10 (i.e., not at all), seems to be overestimated in Q-2 results, especially for NA. In other words, when users state that they do not have NA feelings (lower high NA) in Q-10, they record a higher NA value in Q-2.

Figure 5.b compares the study's mean values of PA and NA per individual. For the majority of the users, the NA is generally overrated in Q-2, when compared to Q-10. Overrating of NA in Q-2 could have various explanations, such as user perception of NA items or limitations of Q-10 items. For example, a user may feel some level of negative activation, but cannot explain it with Q-10 items. In open-ended answers from the post-study survey, a user states that *"I feel negative but not necessarily* hostile *and the negative items cannot show my feelings"*. As a result, users may report some level of negative activation in their self-assessment via Q-2. Furthermore, the PANAS-based scales calculate the sum of all included negative items of the NA dimension. Therefore, the users may feel strongly upset, but the intensity of their feeling per se may not be fully represented by the dimension of NA in the measurement as a whole.

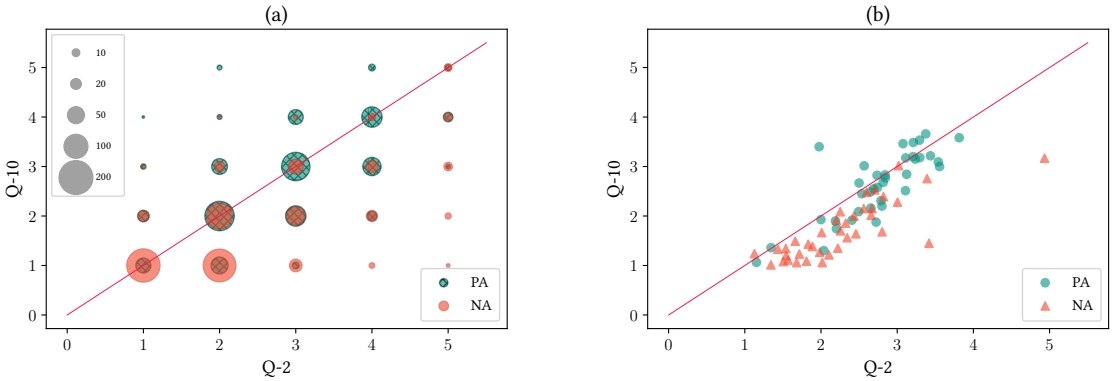

Fig. 5. **(a)** On the left: The distribution of the resulting values of Q-2 and Q-10 for all users in sample A. The bubble sizes represent the frequencies. **(b)** On the right: Comparing the resulting mean values of Q-2 to Q-10 for each individual in sample A. Value 1 was labeled as *not at all* and the value 5 as *a great deal* (see figure 2,). The red line shows identity line.

Stress has a correlation with NA in PANAS in the literature [73]. Q-2, like Q-10, shows an association with stress. We found a significant correlation of $r(765) = .540$, $p < .001$ for NA of Q-10, and, conveniently, a similar correlation $r(765) = .584$, $p < .001$ for NA of Q-2. Watson [73] reported a lower correlation (.44) using PANAS and Benham and Charak [8] reported a higher value (.57) using I-PANAS-SF. Stress, in some studies (e.g., [73]), seems not to have a significant correlation with PA. However, unlike Watson [73], we found a significant negative correlation of $r(765) = -.283$, $p < .001$ between stress and PA, similar to Villodas et al. [71]. This correlation is significantly[9] stronger between PA of Q-2 and stress ($r(765) = -.390$, $p < .001$). But, unlike with NA, this association has not been discussed homogeneously in the literature, and we cannot consider its higher strength as a Q-2 characteristic.

[9]Passes all tests of cocor [22] for comparing the correlations of a dependent group with overlapping variables

The varying correlations in the literature could suggest that stress has stronger correlations with only some items of PANAS. Consequently, the association with stress may not necessarily be captured with all variations of the PANAS measure. In further analysis among Q-10 items, we found stress to have significantly higher correlations with afraid ($r(765) = .514$, $p < .001$) and nervous ($r(765) = .518$, $p < .001$) states, and lower correlations with upset ($r(765) = .381$, $p < .001$), ashamed ($r(765) = .362$, $p < .001$), and hostile ($r(765) = .380$, $p < .001$) states. Among these items, afraid and nervous also have higher loadings in the exploratory factor analysis according to Thompson [69]. The items of a measure seem to influence its correlation with stress. We, therefore, recommend researchers who are interested in stress to carefully choose their measure and the included items, and to consider the limitations and capabilities of the chosen measure with regard to the assessment of stress. Q-2 seems to reflect well on the association of NA and stress, although it does not present any specific mood item of NA.

5.2 Sample B: Weekly Sample

Table 3. The mean, standard deviation, and correlations of the PA and NA values of Q-2 and Q-10 in sample B and C.

	Sample B (n= 258 entries from 129 users)							Sample C (n=177)					
	M	SD	PA-Q-2	PA-Q-10	NA-Q-2	NA-Q-10	Stress	M	SD	PA-Q-2	PA-Q-10	NA-Q-2	NA-Q-10
PA-Q-2	2.832	.827	1					2.643	1.043	1			
PA-Q-10	2.936	.789	**.732****	1				2.490	.896	**.718****	1		
NA-Q-2	2.852	.912	-.517**	-.426**	1			2.729	1.177	-.495**	-.330**	1	
NA-Q-10	2.388	.768	-.319**	-.269**	**.627****	1		1.939	.899	-.360**	-.268**	**.710****	1
Stress	2.043	1.377	-.162**	-.131*	.299**	.329**	1	no assessment was conducted					

The evaluation of sample B revealed results similar to the assessment of sample A. PA and NA items of Q-2 show respectively good internal reliability of *Cronbach's* $\alpha = .881$ and .791 with PA and NA items of Q-10. Table 3 depicts the description of the results for all weekly entries. Although the correlation between PA scores (of Q-2 and Q-10) in sample B (PA: $r(256) = .732$, $p < .001$) is very close (no significant difference) to the resulting correlation in sample A (PA: $r(765) = .723$, $p < .001$), the correlation between NA scores in sample B (NA: $r(256) = .627$, $p < .001$) is significantly lower (using Fisher's Z test: $z = -2.777$, $p = .003$) than sample A (NA: $r(765) = .734$, $p < .001$).

Figure 6 shows the average PA and NA values for each individual in sample B. As with our earlier observation of sample A, participants seem to overrate their NA in Q-2 versus Q-10 (visualized in figure 7, left). In this sample, the intercorrelation between PA and NA values is slightly lower than sample A. NA values resulting from sample B have significantly lower correlations with stress values for both Q-2 and Q-10. One can explain this by considering stress as an intense emotion, rather than a mood state. Stress has physiological responses and fits very well within the definitions of emotion. As a consequence, its effect on mood may decline over time and may not necessarily be reflected in the question that examines the past few days.

5.3 Sample C: One-time Assessment Sample

Q-2 PA and NA items show respectively good internal reliability of *Cronbach's* $\alpha = .886$ and .867 with Q-10 PA and NA items. Table 3 shows the summary of the correlations between the variables in sample C for the default measure, Q-2 followed by Q-10 for all users. Similar to sample A and B, scores of Q-2 have strong correlations with respective scores of Q-10. As a result, it seems that regardless of a one-time assessment or repeated assessment using ESM, Q-2 has a high assessment quality. Comparing Q-2 and Q-10 scores of PA,

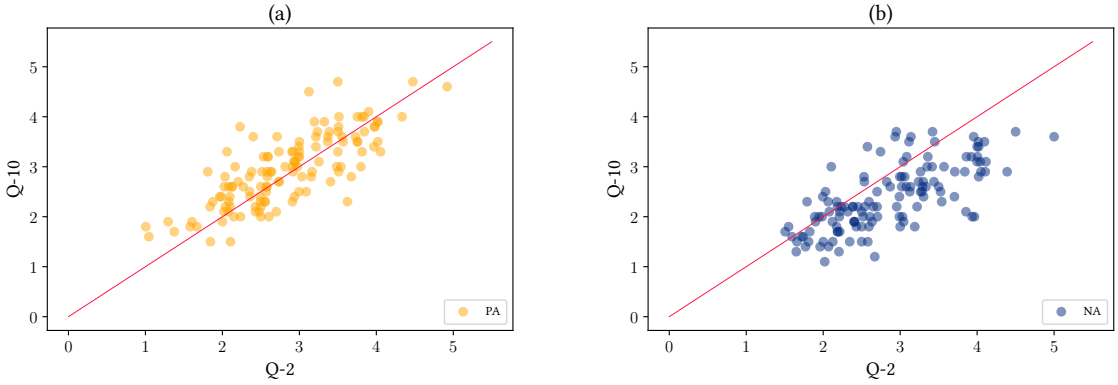

Fig. 6. Comparing the resulting mean values of Q-2 versus Q-10 for each individual in sample B, the weekly mood. Value 1 was labeled as *not at all* and the value 5 as *a great deal*. The red line shows identity line. (a) on the left: PA; (b) on the right: NA

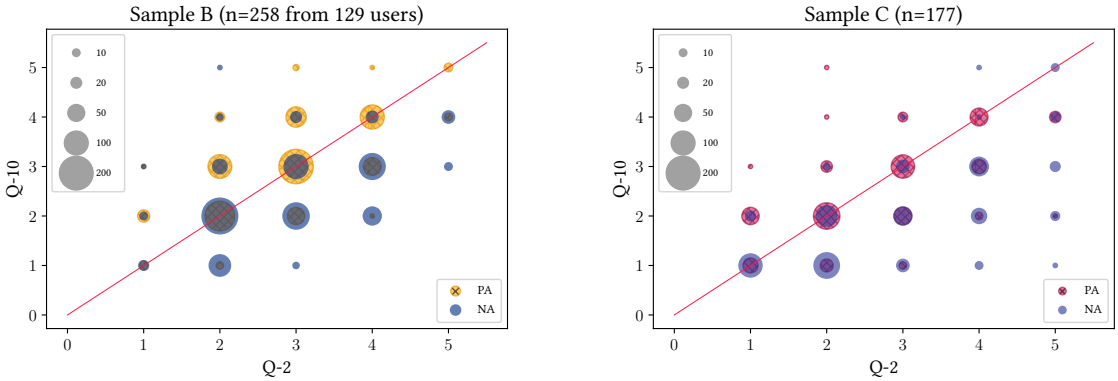

Fig. 7. The distribution of resulting values from the two-item (Q-2) and ten-item (Q-10) measures for all users. Sample B: on the left and C: on the right. Value 1 was labeled as *not at all* and the value 5 as *a great deal*. The red line shows identity line.

resulting from counterbalancing with a chat-based variation, gives a significant correlation of $r(86) = .771$ ($p < .001$) when users answer Q-2 before Q-10, and a correlation of $r(87) = .641$ when users answer Q-2 after Q-10. There is a similar significant correlation of $r(86) = .719$ between Q-2 and Q-10 NA scores when Q-2 is answered before Q-10, and $r(87) = .737$ when Q-2 is answered after Q-10. We used the Cocor package [22] in the R programming language to test the significance of the difference between these two groups (i.e., Q-2 before Q-10 and Q-2 after Q-10). The difference between the two correlations is insignificant testing both Fisher's z [25] and Zou's [80] confidence interval. In other words, the sequence of applying Q-2 and Q-10 (before or after) does not have a significant influence on the resulting correlations.

5.4 Further Discussion

Considering the results from samples A, B, and C, Q-2 items seem to consistently have a high correlation with the resulting PA and NA of Q-10, and the correlation coefficients seem to remain stable with retests. PA of Q-2 generally seems to be stable even in a weekly sample and, therefore, can be reliably used to capture overall PA.

Like Watson et al. [75], we found in all three samples that users record higher values of PA in comparison to NA for Q-10, though it is only the case in sample A for Q-2. Samples A and C capture the mood for *the past few hours* and consistently show a high correlation for NA. However, this correlation seems to attenuate in sample B, which captures the mood for *the past few days*. In sample B, the change in the time-frame of the questions from *hours* to *days* gives a higher value for both PA and NA (reported also by Watson et al. [75]). However, this increase is higher in NA than in PA.

There could be several reasons for the resulting lower strength of correlation in NA compared to PA for sample B (the weekly sample). Overall, as depicted in figure 6, NA captured by Q-2 is higher than that captured by Q-10 on average. It could be because of the broader range and lasting influence of negative affective states that a user may experience within several days, compared to hours. It could also be related to a user's overall memory and recall of those negative experiences. A weekly assessment generally is much less burdensome to users than a daily evaluation. We, therefore, recommend the usage of the full questionnaire (e.g., Q-10) in the weekly assessments despite its moderate correlation with Q-2. Q-2, in particular when the user mood state shows less high negative activation (e.g., not at all), can capture some level of negative activation. User perception of NA items (discussed earlier in the results of sample A for overrating of NA in Q-2), as well as specific memory of negative experiences, could influence such an overrating. Q-2, accordingly, could also give us an overview of user mood state that cannot be represented only with items of Q-10.

Taken together, the stability of the scores over three samples, consistency of the correlations, and good internal consistency between Q-2 and Q-10 items support the assessment quality of Q-2. Q-2 with only two items is short and seems to be a valid and reliable measure to capture a quick overview of PA and NA scores for repeated sampling of mood using smartphones. It gives a good determination of high or low PA and NA values, and, therefore, Q-2 measure seems to be especially useful for mood tracking. Compared to other measures in the community, namely PAM [50], Q-2 clearly reflects the overall mood with PA and NA dimensions better and has a higher assessment quality. Q-2 returns scores that correlate strongly with PA and NA, whereas PAM returns a score that correlates strongly only with PA. Nevertheless, Q-2 measure has limitations in detecting particular affective states. Only a full-item questionnaire can determine specifics of an affective state, and, therefore, should be used for this purpose. In section 7, we will follow the results of the adaptive design and investigate if it could benefit from both Q-2 and Q-10 measures to capture both overall mood and specific mood states.

6 RQ2: LEARNING EFFECT

Results of RQ1 revealed that Q-2 could be used to capture overall mood. However, it remains to be seen if the repeated use of a short measure like Q-2 followed by Q-10 in our data samples impacts its assessment quality or app quality. We hypothesized that, in the beginning, Q-2 might not accurately reflect the Q-10 values of PA and NA, since it could be harder for the users to answer Q-2 questions before becoming familiar with their concepts. Although Q-2 includes a help button, the concepts of positive activation and negative activations could still be ambiguous for users. However, since PA and NA items of Q-10 were presented in two separate pages with a title on top (figure 2.a), users were able to become more familiarized with these concepts as they used the measures (particularly after completing at least three mood entries). Investigating this possibility was the main reason for grouping PA and NA items of Q-10 in separate tabs.

6.1 Method

To address the effect of repetition and user learning on mood sampling (RQ2), we first defined two groups from entries of sample A, with an equal number of entries (87 entries each): one before training (i.e. with a little practice or low level of prior experience with the measures), and one after training (i.e. with more practice or sufficient level of prior experience with the measures). Accordingly, from a total of 174 entries from 29 users,

capturing the first six completed entries per user, the before-training group contains the first three (first, second, and third) completed daily mood entries of each participant. We categorized the next three (fourth, fifth, and sixth) completed entries as the after-training group, in which participants had some practice with and prior experience using the app and were possibly more familiar with the meanings of the terms PA and NA, or with the items to which they refer. With a within-subjects study design, we then examined the influence of user practice (training) on the accuracy of the captured Q-2 values. We then expanded our evaluation to all entries of samples A and B, regardless of the number of completed entries per user.

In addition to the captured variables of PA and NA from both Q-2 and Q-10, we calculated two time-related variables to assess the effort required to complete a mood entry. The first variable, the *response-duration* in seconds, indicates the time that users take to complete and submit the mood entry. The second variable, *response-delay* in seconds, is calculated as the time between the pre-scheduled time of mood sampling (app pop-up) and the time users engage with the app to answer the questions. In this paper, we considered response-duration and response-delay as indicators of the user engagement and the required effort.

Table 4. The learning effect: Comparing mean, standard deviation, and correlations of the PA and NA values across Q-2 and Q-10 measures. Before-training contains the first three completed mood entries per participants whereas after-training includes the next three completed entries when the users had some familiarity with the items and terms of PA and NA. ** Correlation is significant at the P<.001 level.

	Before Training							After Training						
	Mean	SD	Mdn	1	2	3	4	Mean	SD	Mdn	1	2	3	4
1. PA: Q-2	2.734	.913	2.930	1				2.784	1.011	2.980	1			
2. PA: Q-10	2.630	.836	2.400	**.798****	1			2.754	.981	2.800	**.741****	1		
3. NA: Q-2	2.321	1.156	2.000	-.554**	-.391**	1		2.469	1.120	2.200	-.601**	-.423**	1	
4. NA: Q-10	1.685	.815	1.400	-.405**	.198	**.708****	1	1.807	.875	1.400	-.457**	-.285**	**.812****	1
Stress	1.962	1.384	1.980	-.288**	.180	.518**	.485**	2.235	1.471	2.160	-.469**	-.359**	.670**	.529**
Response-duration	≈10 min	≈33 min	49.263 sec					≈9 min	≈20 min	52.224 sec				
Response-delay	≈27 min	≈78 min	0.408 sec					≈45 min	≈152 min	0.375 sec				
n	87													

6.2 Results and Discussion

Table 4 shows the correlation coefficients, mean, and standard deviation of the variables comparing before-training and after-training groups. Assessing the two groups did not reveal any significant difference between them for any of the captured variables. This could be due to the definition of the groups, which limit them to a total of only six completed entries per user. Learning by practice may have a gradual and subjective effect. To assess this further, we extended our evaluation by ignoring the assumed condition of before-training and after-training groups, and instead considered all entries of sample A.

However, users of sample A had varying numbers of entries. We therefore, in addition to all entries of sample A, also looked into a filtered sub-sample consisting only of 27 users with an equal number of completed entries (12 successful repetitions for each user). Figure 8 (a, b) visualizes the correlations of Q-2 and Q-10 per sequence of completed entries (i.e., successful repetition). The left diagram (a) shows the filtered sub-sample. In contrast, the right diagram (b) includes entries from all users that have a varying number of completed entries (illustrated by the size of the points in figure 8.b). With each sequence of completed entries (i.e., repetition), the correlations between Q-2 and Q-10 values fluctuate; nonetheless, this fluctuation seems inconclusive with regard to the number of repetitions. The response duration does not seem to change with more practice (figure 8 (c, d)). The average response duration is in the range of 4.68 to 27.12 minutes[10]. Although the second sequence in figure 8.c

[10]For the mood entries that were captured before the 25th completed entry

initially suggests a reduction in the mean response duration, other factors — rather than only the number of times a user answers the measure — seem to influence the response duration. The same applies to the response-delay.

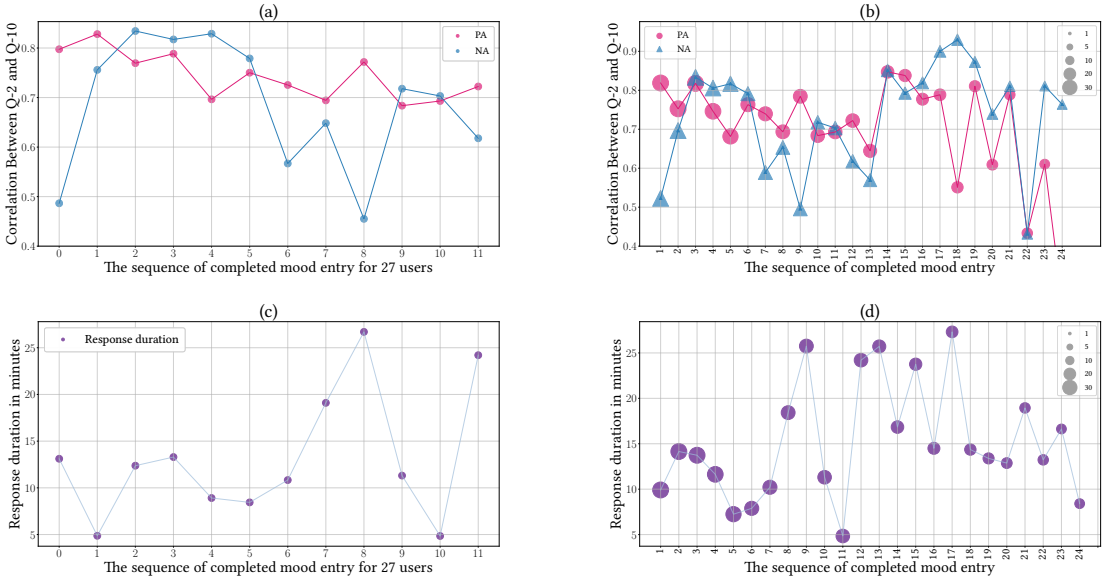

Fig. 8. **(a, b)**: The correlation between Q-2 and Q-10 for each sequence of completed mood entry for **(a)**: 27 users who have completed all 12 sequences; **(b)**: all individuals. **(c, d)**: Average response-duration in minutes per sequence for **(c)**: 27 users with 12 completed sequences each; **(d)**: all individuals. The circle/triangle size in **b** and **d** depicts the number of users.

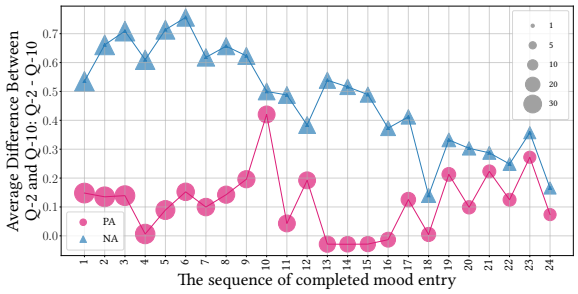

Fig. 9. The mean differences between Q-2 and Q-10 values for each individual. The closer the difference to zero, the closer the resulting values of Q-2 and Q-10.

The subtraction of Q-10 scores from Q-2 scores, however, seems to suggest a change over repeated successful use of the measure. Figure 9 shows the average of the difference between Q-2 and Q-10 scores for each sequence. Over time, with more completed entries, the average difference between the Q-2 and Q-10 scores generally gets closer to zero (more apparent for NA). It seems that PA item are more intuitive than NA item, and user perception

of NA item is not as accurate as of PA item. For this reason, NA accuracy slightly improves with repetition or practice. Nevertheless, this difference between sequences (the number of repetitions) is not statistically significant for any calculated or measured variables.

We also evaluated the weekly sample B. Despite slight variations in the correlation coefficients per sequence of successful interaction with the weekly measure, there is no significant difference between the sequences as the interaction increases. In the end, we found no definite evidence on the effect of the repeated use of the measure (on a daily or weekly-basis (sample A and B)) on user learning and, accordingly, providing a significantly more accurate determination of overall mood using Q-2. Repeated interaction with the measure did not reveal any significant improvement in the response duration or delay. In other words, repeated interaction and learning seem to influence neither the assessment quality nor the required effort. This could especially be the case for non-complex measures that are easy to use and self-explanatory (as indicated by users in the usability assessment of the measure, see section 7). As a result, one should not be concerned with the learning effect while using such measures repeatedly, concerning the discussed variables. One should still keep in mind that the learning effect in complex measures, such as Affect Grid [57], could potentially lead to a different outcome.

7 RQ3: ADAPTIVE DESIGN

While Q-2 is a valid and short measure for capturing overall mood, it does not provide any details. For example, if there were a major increase in the overall NA value, Q-2 alone, could not indicate which mood states caused this increase. In order to have a short measure and capture details of a user's state, we here explore the design and functionality of a basic adaptive measure, and answer RQ3. We also investigate whether adaptively reducing the length of a mood measure would impact its usability and user compliance.

7.1 Method

7.1.1 Adaptive Measure. The adaptive design used here is based on the classic I-PANAS-SF scale as described in Q-10 (see 3.3). However, instead of always showing Q-10 right after Q-2 like sample A, we show only a fraction of Q-10, based on the values of Q-2. The idea of this adaptive measure was inspired by the initial results of the pre-study (see 4.3), which suggested that the Q-2 values highly correlate with values of Q-10; therefore, we could potentially predict if a fluctuation exists by using a basic algorithm. These correlations were later confirmed by ESM (see 4.1) and weekly assessments (see 4.2) as well (see table 2 and 3 in section 5). Predicting fluctuations would help to determine a user's current mood quicker and with the minimum number of questions, and it could be more comfortable for users to record their mood. The adaptive design algorithm would run in every instance of mood sampling for each user of adaptive measure.

The logic of this adaptive design (illustrated in algorithm 1) aims to reduce the number of questions in each measurement depending on the variation of the overall score. If a user is in the same mood state as before (i.e., no fluctuation occurs), no additional questions would be asked, and we could assume that the user is still in the same mood state. This adaptive design also needs to consider the temporal nature of mood while using preceding records of mood. In order to achieve these targets, the algorithm first uses a default threshold value of one scale unit that represents the subjective sensitivity of individuals when entering their mood using a Likert-type scale. Taking the last four completed measures that are recorded within the maximum of the past two days, it calculates the mean and standard deviation of these measures. Accordingly, it updates the threshold values of PA and NA, with the resulting standard deviation values, if they are less than the default threshold value.

In this adaptive measure, users see the Q-2 questionnaire and, based on their PA and NA values of Q-2, the algorithm determines if there is a mood fluctuation. When the Q-2 values are not between mean, plus and minus the threshold values, it is assumed that there is a mood fluctuation. A fluctuation is also assumed if Q-2 values are not between the values of PA and NA from the previous measurement, plus and minus the standard deviation.

Algorithm 1: The logic of the adaptive design

Input: A finite set $X = \{x_1, x_2, \ldots, x_{n-1}\}$ of either PA (or NA) values sorted by order of entry. Values of incomplete (invalid or dismissed) entries are null.
Output: Whether to show the detailed items of Q-10 for only PA, only NA, both, or neither.

$shouldPresentDialogue \leftarrow True$;
$threshold \leftarrow 1.0$;
x_n =value of PA (or NA) from Q-2;
if $x_{n-4} = null \vee x_{n-3} = null \vee x_{n-2} = null \vee x_{n-1} = null \vee (TSTP_{x_n} - TSTP_{x_{n-4}}) > twoDays$ **then**
    return *shouldPresentDialogue*;
else
    $mean \leftarrow Mean(x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1})$;
    $std \leftarrow STD(x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1})$;
    if $std < threshold$ **then**
        $threshold \leftarrow std$;
    if $(x_n < mean + threshold) \wedge (x_n > mean - threshold)$ **then**
        $shouldPresentDialogue \leftarrow False$;
    if $(x_n < x_{n-1} + std) \wedge (x_n > x_{n-1} - std)$ **then**
        $shouldPresentDialogue \leftarrow False$;
    return *shouldPresentDialogue*;

Conversely, if Q-2 values do not meet any of the mentioned conditions, we expect that no mood fluctuation has taken place. If any of the last four measures are not present (i.e., invalid or dismissed), we infer that we cannot accurately determine the mood without recent historical data. Subsequently, the Q-10 questionnaire with all items would be shown to the user (i.e., the fluctuation is assumed). Based on both the result of the algorithm and the resulting assumption of fluctuation, more detailed questions of either PA, NA, or both from Q-10 would be shown to the users. In other words, depending on their response to Q-2, users would either see zero questions, five PA questions, five NA questions, or all ten PA and NA questions of Q-10. The final score of PA and NA for a sampling event would be either recorded directly from Q-2 or calculated from the shown Q-10 items.

7.1.2 Procedure. To investigate RQ3, we first studied the validity and reliability of the adaptive design algorithm with an offline evaluation of the collected entries of sample A. We also looked into sample D, which is the application of the adaptive design, using a twice-daily ESM. Participants of sample A (section 4.1) always saw Q-2 followed by Q-10, while participants of sample D (section 4.4) always saw Q-2 followed by either a partial set, a complete set, or none of the Q-10 items. All conditions of the study, except the measure used for both samples of A and D, were the same. We used exactly the same dialogues and interface for both measures in samples A and D. Finally, we compared sample D with sample A (i.e. a between-subjects study design) over the defined period of study (two weeks) concerning the usability evaluation of the measure, user experience, user compliance, response rate, and response-delay. The user experience of measures used by participants of sample A and D, were captured via the post-study survey functionality of the app (explained in 3.4).

7.1.3 Post-study Survey: User Experience of Measures. The post-study survey focuses on the usability evaluation of the app and, correspondingly, its mood measures. Users followed the survey's link after activation to access the survey page online, using either desktop or smartphone. Users of samples A and D both received the link to the post-study from the app. Overall, 19 users of the adaptive mood measure, and 16 users of the Q-2/Q-10 version,

filled out the survey. We removed those entries in which users claimed they often had submitted meaningless responses. We also limited the survey responses to those participants that were included in samples A and D (i.e., passed the inclusion criteria test of the study section 3.5). This left 12 users from sample A and 13 users from sample D, who completed the post-study survey. The survey has the following elements:

(i) Briefing, debriefing, and reluctance and meaningless response questions
(ii) Modified questions on the mood tracker app asking about the app completion, future-use, effectiveness, emotional gain, usefulness, and stimulation (internal reliability Cronbach's $\alpha = .896$), and ease of use, time-consumption, annoyance, and learnability (internal reliability Cronbach's $\alpha = .743$)
(iii) Open-ended questions on what users liked and disliked about the app
(iv) User Experience Questionnaire (UEQ) [39] (internal reliability Cronbach's $\alpha = .948$)
(v) System Usability Scale (SUS) [10, 11] (internal reliability Cronbach's $\alpha = .963$)

7.2 Results and Discussion

Evaluation of samples A and D shows that the adaptive measure reduces the number of questions between about 30%–60%. We applied the adaptive algorithm to sample A and compared the captured values of Q-10 to the predicted values from the algorithm in an offline evaluation. This led to 226 cases (29.46%) of skipping the Q-10 questions (assumption of no fluctuation). The correlation between the Q-2 score and the captured Q-10 value in these 226 entries for PA is $r(166) = .799, p < .001$ and for NA is $r(125) = .825, p < .001$. The mean absolute errors for PA and NA, respectively, are $.444$ and $.404$, and the accuracies are 60.71%, and 56.69%. The adaptive algorithm acts slightly better than random. Reducing the number of preceding entries included in the algorithm would decrease the number of skipped questions for both PA and NA. It would also reduce the accuracy of PA estimation, but increase the accuracy of NA estimation, between about 6%–9% for both PA and NA. The number of skipped Q-10 questions, consistently would be higher for PA than NA, as well. The striking observation here is the difference between PA and NA with regard to the preceding mood states, which suggests that the modeling and prediction of PA and NA should be made possibly differently.

Analysis of sample D reveals that Q-10 questions were skipped for 697 instances out of $1,080$ completed entries (i.e., 64.5% in practice[11]). Accordingly, users in sample D answered 62.3% fewer questions in total. Samples A and D have, respectively, a per user average of $M = 39.51, SD = 15.56$, and $M = 39.69, SD = 16.738$ entries (*completed*, *dismissed*, and *invalid*). Considering the similar number of entries, it seems that users of sample D ($M = 22.5, SD = 13.99$) completed more mood samplings in comparison to users of sample A ($M = 20.73, SD = 14.05$). Despite the considerable difference between the total number of questions that users in sample D had to answer ($8,080$ questions) and the number of questions they would have had to answer without the adaptive algorithm ($12,960$) – as users of sample A did – none of the variables comparing sample A and D, including the post-survey usability evaluation, show any significant difference.

When measuring variables, such as response-delay or completion rate, one would expect to see a difference in a longitudinal study. However, we observed only a slight increase in the overall completion rate for the period of two weeks. This suggests that user compliance may not be influenced by the length of the questions as much as expected. We can observe a more complete picture when considering the whole population of the study without limitations (i.e., all users before applying the study inclusion criteria described in section 3.5). Sample A without limitations includes 90 users, and sample D without limitations has 92 users. Both are randomly assigned. No significant difference between the two unlimited samples for any of the measured variables was found. Nevertheless, the major difference may reveal itself in the number of users in each sample. Applying the same study condition to both samples A and D results in 37 users in sample A and 48 users in sample D. One may think that the number of questions or the length of a measure would influence the number of dropouts at

[11] 219 cases only PA skipped, 199 only NA skipped, and 279 entries have both PA and NA skipped

the beginning of the study, considering that the dropout rate of 59% in sample A is higher than 48% of sample D. Nevertheless, this assumption would not apply here, since sample D requires at least two days of measurement to reduce the number of questions, and most of the dropouts from both samples occur within the first three days.

The post-study survey gives us an overview of the mood measures' usability. The measure of Q-10 (used in sample A) results in an average SUS score of $M = 77.91, SD = 15.03, Mdn = 80$, suggesting very good usability. This score shows high usability for Q-10 when compared to other mood measures designed for smartphones, such as [27]. The average SUS score of the adaptive measure (used in sample D) is $M = 83.65, SD = 13.60, Mdn = 85$, which suggests excellent usability, according to Bangor et al. [5]'s definition. As figure 10 depicts, users of sample D, on average, gave a higher SUS score to their measure when compared to users of sample A. The average score of the adaptive measure is slightly higher than Q-10 for all items of SUS, except items of complexity and well-integrated functions (figure 11.a). A similar result emerges for UEQ. Figure 11.b depicts the mean score of all factors in UEQ. Although the difference of the scores between the measures is not significant, all mean values are higher in adaptive measures when compared to Q-10. This is especially interesting because both measures have the same design and user interface, the only difference being the algorithm behind the adaptive measure. Therefore, slightest modifications can potentially impact the app quality. Nevertheless, although adaptive behavior has consistently better average SUS and UEQ scores than Q-10, it resulted in no significant improvement of the usability scores, possibly due to the small sample size of the post-study survey.

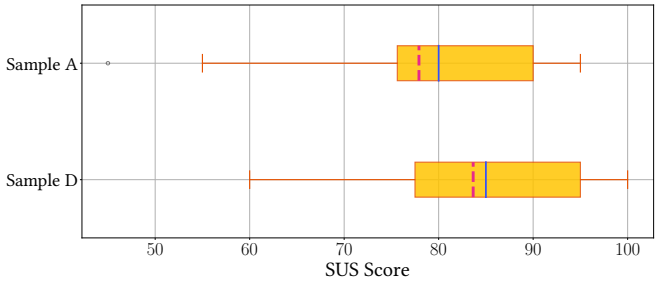

Fig. 10. Resulting SUS score of the two measures of sample A and sample D. The pink dashed line represents the mean and the blue straight line is the median.

The observed consistent difference between the perceived usability of the two measures could be related to the user expectation and preference toward answering fewer questions. Users of the adaptive measure generally, at the time of interaction, do not know whether they will receive a full or partial questionnaire. Their decision to engage with a sampling event is mainly made before engaging with the measure (the number of invalid entries was very small compared to dismissed or completed in all four samples, meaning that, when users start answering, it would be very unlikely for them to quit at the middle, e.g., due to the length of the measure). Consequently, this decision may not be influenced by the length of a measure, since there is no consistent or predictable set of questions in each sampling, and the questions are adaptively reduced. For this reason, there is no observable difference in the response rate, user compliance, response-delay, or dropouts between the measures. However, since users encounter a shorter measure with fewer questions in total, it may improve their experience with a measure or even engage their anticipation throughout the study, leading to slightly better UEQ and SUS scores. Because users of Q-10 measure have to answer more questions, they may be generally bored. As a result, they feel more dissatisfaction toward the app usage experience than users of the adaptive measure. However, further studies should explore other possibilities for the difference between the measures concerning user expectations and usability.

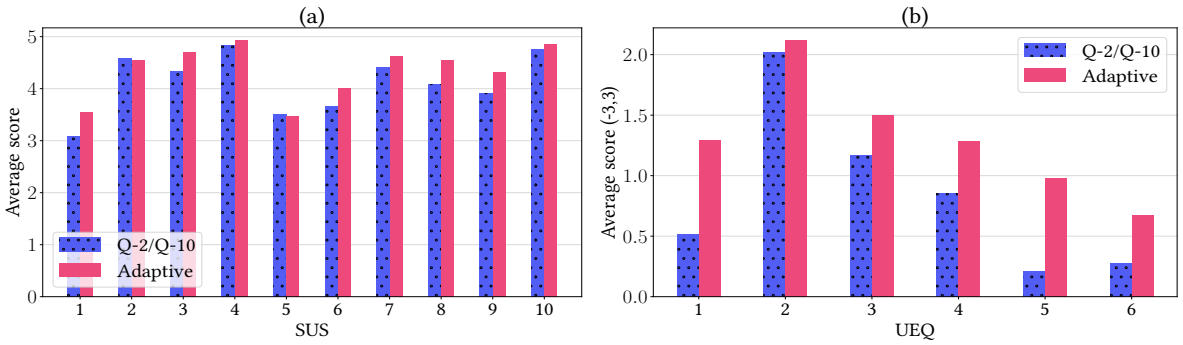

Fig. 11. **(a)** SUS items, average per measure. (R): are negative items and are depicted in the diagram in the reverse form. X-axis: 1. Frequent use, 2. Complexity (R), 3. Ease of use, 4. Need technical support (R), 5. Well-integrated functions, 6. Inconsistency (R), 7. Learnability, 8. Cubersome to use (R), 9. Confident using, and 10. Have to learn a lot (R). **(b)** UEQ items, average per measure. X-axis: 1. Attractiveness, 2. Perspicuity, 3. Efficiency, 4. Dependability, 5. Stimulation, and 6. Novelty.

Comparing the adaptive and Q-10 measures provides a baseline for further advanced machine-learning or time-series algorithms used to predict the mood state from previous records of an individual. It also produces a baseline for the usability of such measures. Newly designed measures may claim to provide a better usability or user experience only if they outperform the classic measures, such as Q-10. This study only discussed the usability and compliance of two measures that were essentially identical. As seen above, the slightest modifications, even in the algorithm of the measure, can influence the resulting usability in a longitudinal study, despite having the same visual design or interaction. Therefore, usability assessment and claims of a measure should be carefully considered and compared.

An adaptive measure can reduce the total number of questions up to 60% for the duration of a longitudinal study. It can also capture mood fluctuation using Q-2 and provide more details on specific user mood states. A smaller variable set of questions does not necessarily lead to significantly higher or better compliance, fewer dropouts, response-delay, and usability. It may, however, slightly improve the user experience due to variation in questions, engagement of user anticipation, or fulfillment of user expectations to receive fewer questions. It can also increase the completion rate, though very slightly. Other factors, such as the usability of the design, interruptability of users throughout the day, or user context, might play a more dominant role in user compliance. As a result, it is advisable to consider the application of the full measures instead of an adaptive solution, where applicable.

8 FURTHER DISCUSSION, LIMITATIONS, AND FUTURE WORKS

This paper has gone some way towards balancing app quality and assessment quality and increasing our understanding of mood tracking. We looked into the assessment quality of overall questions, instead of the lengthier PANAS. We then investigated the feasibility and potentials of an adaptive measure, and we compared its app quality with the lengthier classic measure, which can be used as the app quality baseline for further comparisons. We found out that Q-2, (i.e. overall questions of PA and NA), is reliable and valid for quickly capturing mood, and has a high assessment quality. We also learned that the repeated use of Q-2 followed by Q-10, does not significantly change its assessment quality nor app quality (or, in this case, the learning by practice that can be reflected by quicker response time, response delay, or closer scores of Q-2 and Q-10). Adaptively reducing the length of a mood measure can reduce the total ESM questions by up to 60%; however, such a reduction may not significantly improve the app quality, which may be due either to user expectations at the time of engagement

or to the limitations of our study. A remarkable result to emerge from the data was the difference between PA and NA concepts. Based on the evidence from our results, the concept of PA seems to be more self-explanatory than NA, since the average difference between NA scores of Q-2 and Q-10 gets closer, although insignificantly, to zero with repeated use of the measures. In contrast, the average difference between PA scores of Q-2 and Q-10 seems to be stable over time. We also discovered that NA is either longer-lasting or easier to recall than PA to an extent which would impact the value of overall PA and NA, and, therefore, should be modeled differently.

These observations have several implications for research into well-being and building mood-aware systems, as well as for the interdisciplinary use of all questionnaire-based measures. Our results encourage the scientific community to avoid using arbitrary measures, and to treat assessment quality with the utmost caution. They also demonstrate that even lengthy classic measures of mood can have good app quality, from the user's perspective. Therefore, the contribution related to improvements in app quality should be cautiously confirmed, with regards to classic measures as the baselines. In our view, the results of Q-2 and adaptive measures constitute an excellent initial step toward mood tracking with both high assessment quality and app quality. They also maintain that using overall questions from a well-established theoretical construct is better than many other devised measures that are not compatible with earlier theories of affect, mood, and emotion.

Our results of Q-2 are in line with the classic measures of mood such as PANAS or I-PANAS-SF; nevertheless, it is plausible that a number of limitations could have influenced the results obtained. To investigate the learning effect in RQ2, we intentionally grouped PA and NA items in Q-10. Although some studies in other domains, such as personality and emotional stability [58], suggest that grouping items may not significantly impact the assessment quality, this could have potentially influenced a mood scale like I-PANAS-SF. Nevertheless, our results regarding the intercorrelation between PA and NA using Q-10 is consistent with and similar to previous results using the original forms of PANAS and I-PANAS-SF [17, 26, 60, 68, 75] (discussed in section 5.1). We can therefore assume that grouping PA and NA items did not impact the assessment quality of Q-10 in our study. However, we recommend researchers to randomize these items where possible. Q-2 seems to have high discriminant correlations when compared to I-PANAS-SF (i.e. Q-10), which can be related to the orientation of PA and NA items in one dialogue [78]. However, it could also be related to the reduced number of questions. On the one hand, reducing the number of PANAS items in I-PANAS-SF increased the intercorrelation between PA and NA scores, according to Thompson [69]. On the other hand, researchers have reported higher intercorrelation than what we found for Q-2, even when using the original PANAS measure [26], and particularly when considering measurement errors. As a result, although the intercorrelation of PA and NA in Q-2 corroborates earlier findings, the discriminant correlations of this measure should be further explored.

The ESM app had a feedback functionality, which showed the users' mood of the day, weekday moods, and mood fluctuations. This functionality was integrated into the app as a way to engage participants with the ESM study, according to Hsieh et al. [28]. One may think that the user feedback could potentially have some influence on the accuracy of self-reported mood, since a user may intentionally moderate their self-reported responses, depending on the received feedback. This possibility should be further investigated. Nonetheless, even if such a limitation exists, it would apply similarly to any self-reported mood tracking, including classic measures and measures of this paper. In other words, using a feedback mechanism with mood tracking and its effects on response veracity is a limitation related to all self-reported measures. Among our data samples, only independent samples A, B, and D were ESM-based, and their users saw visualized feedback, whereas sample C was from a one-time pre-study assessment. Nevertheless, we did not observe any noticeable difference in the average recorded mood between samples or any evidence from the retrieved data, suggesting that the self-reported responses are moderated because of the feedback functionality.

To investigate user effort and engagement in this paper, we looked mainly into response duration and response delay variables. However, additional variables such as user compliance, user endurance, completion rate, and dropout rates, could draw a more complete picture of user effort and engagement. In this paper, using ESM, we

wanted to determine the assessment quality of Q-2, and therefore, we had to have a concurrent within-subjects assessment of Q-10 together with Q-2. As a consequence, user compliance and other variables would not fully reflect Q-2 in our samples, unless a user uses only Q-2. In general, a two-item measure should be easy to interact with and quick to use and should potentially have a better compliance rate than longer measures. Knowing the validity of Q-2, one can later compare it with Q-10 or other measures as an independent measure with a between-subjects ESM study, and further, varify the improvement and measure its extent for other variables.

Our participants were recruited through study announcements and media promotions, and participants did not receive any financial compensations for their participation. Nevertheless, the overall response-rate and user compliance in our ESM studies were consistent with previous ESM studies in the community [30, 70]. In their meta-analysis, van Berkel et al. [70] reported a lower response rate for raffle-based or no compensation methods (on average less than 60%). Intille et al. [30] reported a user sample of 38 remaining participants from 88 initial volunteers, and they found an overall compliance rate of 53.28% for the first ESM prompt, which is similar to our ESM event, since we prompted an event only once. In sample A and D (our twice-daily ESM), we have 38 and 48 users remaining from an initial 90 and 92 interested individuals, respectively. The overall compliance rate, calculated in a manner similar to [30] for sample A, is around 52%, and for sample D is about 56%. Both measures of sample A and D had a cancel button on the interface, and by using it, users could easily dismiss the sampling event. Considering that finishing the measurement in our studies would require far more user effort than dismissing it (with a total of 13 consecutive questions from Q-2, Q-10, and stress measure), it would be unlikely for a user to submit meaningless responses instead of dismissing the events. This was also confirmed by assessing the response duration of the completed measure that was within a reasonable range. Normally, in addition to app quality, user motivation can affect user compliance in an ESM study. This would not, however, severely impact the contributions of this paper, since the user recruitment procedure for all study conditions (randomly assigned) and subsequently, data samples of this paper, were the same.

To evaluate the app quality, we aimed to complete a thorough examination of the measures of this paper. The user experience and usability evaluation of an app is usually based only on first impressions. Users see at most an interface or prototype, or interact once with the app, and then answer a survey or interview questions and indicate their opinions. Despite the potentials of such evaluations, a one-time assessment or first impression does not reveal aspects related to repeated use or ESM, which is particularly important in the app quality of mood tracking. Given the focus of this paper in addressing the app quality of mood tracking, we used the post-study survey to capture the users' evaluations of the app, and specifically designed the study in such a way that only after two weeks of participation, users were able to access the survey. Inevitably, the number of users who completed the survey was limited. Furthermore, using the measure for two weeks limits the study to a between-subjects design, which has its own limitations. For example, users may not have experience using any other instrument with which to compare the measure as a baseline, or they may just be critical of the app or measure. Tests carried out with Q-10 confirmed that this classic measure has very good usability from the user's perspective. They also revealed that the user experience of a classic self-explanatory measure, such as Q-10, would not be very different from other measures in between-subjects evaluations. However, the main difference would possibly reveal itself in user compliance and endurance of the study. Therefore, further studies need to be undertaken to investigate both user compliance and endurance in an ESM study, as well as side-by-side within-subjects comparison of the app quality.

9 CONCLUSION

This paper discussed the application of smartphone-based mood measures in longitudinal studies and the challenges that a researcher could face with the modification or usage of classic measures, such as PANAS. We specifically focused on approaches to make a measure shorter and easier to administer using smartphones,

and conducted several studies to explore our research questions. By providing a comprehensive evaluation of a shorter version of PANAS, I-PANAS-SF, together with a two-item-based overall self-assessment of PA and NA, we found that users are capable of providing a general assessment of their mood using the overall questions. The resulting two-item-based measure, Q-2, is a valid and reliable measure for capturing mood fluctuation over time. The resulting NA is associated with user stress levels in daily assessments. We found that researchers using non-complex and self-explanatory measures should not concern themselves with learning effects in the repeated use of a measure regarding response duration, delay, or the quality of the responses.

We designed an adaptive measure that, based on the mood fluctuation, asks only specific questions and reduces the number of questions a user has to answer. While investigating this measure, we found that even the slightest modifications can change a user's experience of a measure. Therefore, researchers need to be cautious of the slightest design variations or modifications of a measure in smartphones. We also found a baseline for more advanced prediction models, as well as further usability improvement of the measures in the future. We found that an adaptive reduction of the questions in a measure may not necessarily lead to higher compliance, better response rate, or significantly better usability. Most of the dropouts occur within the first two days of a study. As a result, users are likely to continue interacting with easy-to-use measures independent of the total number of questions they have to answer in the end. Considering the results of this study, we would like to further focus on the question of compliance and interaction design in the future by comparing various measures with different designs. It would also be worthwhile to investigate the benefits of the measure's length in a stable design (e.g., two-item-based, or complex short measures) compared to the Q-10 measure. By excluding the adaptiveness of a measure and comparing two stable and consistent measures, we can explore other possible influences a measure's length has on user compliance and experience.

ACKNOWLEDGMENTS

This research is partially funded by the German Federal Ministry of Education and Research (BMBF) as part of the project PAnalytics (16SV7110).

REFERENCES

[1] [n.d.]. Google Scholar. https://scholar.google.de/scholar?hl=en&as_sdt=0%2C5&q=panas&btnG= accessed on July 2020.
[2] [n.d.]. PubMed. https://pubmed.ncbi.nlm.nih.gov/3397865/ accessed on July 2020.
[3] [n.d.]. Web of Science. http://apps.webofknowledge.com/ accessed on July 2020.
[4] Anja Bachmann, Christoph Klebsattel, Matthias Budde, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. How to Use Smartphones for Less Obtrusive Ambulatory Mood Assessment and Mood Recognition. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 693–702. https://doi.org/10.1145/2800835.2804394
[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4, 3 (May 2009), 114–123. https://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/
[6] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An Inventory for Measuring Depression. *Archives of General Psychiatry* 4, 6 (June 1961), 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004 Publisher: American Medical Association.
[7] Christopher J Beedie, Peter C Terry, and Andrew M Lane. 2005. Distinctions between emotion and mood. *COGNITION AND EMOTION* 19, 6 (2005), 847–878.
[8] Grant Benham and Ruby Charak. 2019. Stress and sleep remain significant predictors of health after controlling for negative affect. *Stress and Health* 35, 1 (2019), 59–68. https://doi.org/10.1002/smi.2840
[9] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49 – 59. https://doi.org/10.1016/0005-7916(94)90063-9
[10] John Brooke. 2013. SUS: a retrospective. *Journal of Usability Studies* 8, 2 (Feb. 2013), 29–40.
[11] John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[12] Rafael A. Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (Jan. 2010), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

[13] Yu-Lin Chang, Yung-Ju Chang, and Chih-Ya Shen. 2019. She is in a Bad Mood Now: Leveraging Peers to Increase Data Quantity via a Chatbot-Based ESM. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, Taipei, Taiwan, 1–6. https://doi.org/10.1145/3338286.3344406

[14] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, Toronto, Ontario, Canada, 1143–1152. https://doi.org/10.1145/2556288.2557372

[15] Lee Anna Clark, David Watson, and Jay Leeka. 1989. Diurnal variation in the Positive Affects. *Motivation and Emotion* 13, 3 (Sept. 1989), 205–234. https://doi.org/10.1007/BF00995536

[16] Marios Constantinides, Jonas Busk, Aleksandar Matic, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2018. Personalized versus Generic Mood Prediction Models in Bipolar Disorder. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. Association for Computing Machinery, Singapore, Singapore, 1700–1707. https://doi.org/10.1145/3267305.3267536

[17] John R. Crawford and Julie D. Henry. 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 (Sept. 2004), 245–265.

[18] Mohamed Dahmane, Jahangir Alam, Pierre-Luc St-Charles, Marc Lalonde, Kevin Heffner, and Samuel Foucher. 2020. A Multimodal Non-Intrusive Stress Monitoring from the Pleasure-Arousal Emotional Dimensions. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.2988455

[19] Victor-Alexandru Darvariu, Laura Convertino, Abhinav Mehrotra, and Mirco Musolesi. 2020. Quantifying the Relationships between Everyday Objects and Emotional States through Deep Learning Based Image Analysis Using Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 7:1–7:21. https://doi.org/10.1145/3380997

[20] Charles Darwin. 1872. *The expression of the emotions in man and animals*. Vol. 526. University of Chicago press, London.

[21] P. M. A. Desmet, M. H. Vastenburg, and N. Romero. 2016. Mood measurement with Pick-A-Mood: Review of current methods and design of a pictorial self-report scale. *Journal of Design Research* 14, 3 (2016), 241–279. https://doi.org/10.1504/JDR.2016.079751 Publisher: Inderscience Enterprises Ltd.

[22] Birk Diedenhofen and Jochen Musch. 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE* 10, 4 (April 2015), e0121945. https://doi.org/10.1371/journal.pone.0121945 Publisher: Public Library of Science.

[23] Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion* 3 (1984), 19–344.

[24] Angela Fessl, Verónica Rivera-Pelayo, Viktoria Pammer, and Simone Braun. 2012. Mood Tracking in Virtual Meetings. In *21st Century Learning for 21st Century Skills (Lecture Notes in Computer Science)*, Andrew Ravenscroft, Stefanie Lindstaedt, Carlos Delgado Kloos, and Davinia Hernández-Leo (Eds.). Springer, Berlin, Heidelberg, 377–382. https://doi.org/10.1007/978-3-642-33263-0_30

[25] Ronald A. Fisher. 1952. Statistical Methods for Research Workers. https://psychclassics.yorku.ca/Fisher/Methods/index.htm

[26] Donald P. Green, Susan L. Goldman, and Peter Salovey. 1993. Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology* 64, 6 (1993), 1029–1041. https://doi.org/10.1037/0022-3514.64.6.1029

[27] Pegah Hafiz, Raju Maharjan, and Devender Kumar. 2018. Usability of a mood assessment smartphone prototype based on humor appreciation. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. Association for Computing Machinery, Barcelona, Spain, 151–157. https://doi.org/10.1145/3236112.3236134

[28] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E. Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th international conference on Ubiquitous computing*. Association for Computing Machinery, New York, NY, USA, 164–167. https://doi.org/10.1145/1409635.1409657

[29] Kent E. Hutchison, Robert P. Trombley, Frank L. Collins, Daniel W. McNeil, Cynthia L. Turk, Leslie E. Carter, Barry J. Ries, and Michael J. T. Leftwich. 1996. A comparison of two models of emotion: Can measurement of emotion based on one model be used to make inferences about the other? *Personality and Individual Differences* 21, 5 (Nov. 1996), 785–789. https://doi.org/10.1016/0191-8869(96)00107-9

[30] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, Heidelberg, Germany, 1124–1128. https://doi.org/10.1145/2971648.2971717

[31] Shoya Ishimaru and Koichi Kise. 2015. Quantifying the Mental State on the Basis of Physical and Social Activities. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1217–1220. https://doi.org/10.1145/2800835.2807934

[32] Carroll E Izard. 2010. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review* 2, 4 (2010), 363–370.

[33] William James. 1884. What is an emotion? *Mind* os-IX, 34 (1884), 188–205.

[34] Jahanvash Karim, Robert Weisz, and Shafiq Ur Rehman. 2011. International positive and negative affect schedule short-form (I-PANAS-SF): Testing for factorial invariance across cultures. *Procedia - Social and Behavioral Sciences* 15 (Jan. 2011), 2016–2022. https://doi.org/10.1016/j.sbspro.2011.04.046

[35] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 629–641. https://doi.org/10.1145/3025453.3025750

[36] Kyle Kercher. 1992. Assessing Subjective Well-Being in the Old-Old: The PANAS as a Measure of Orthogonal Dimensions of Positive and Negative Affect. *Research on Aging* 14, 2 (June 1992), 131–168. https://doi.org/10.1177/0164027592142001

[37] Le Minh Khue, Eng Lieh Ouh, and Stan Jarzabek. 2015. Mood self-assessment on smartphones. In *Proceedings of the conference on Wireless Health (WH '15)*. Association for Computing Machinery, Bethesda, Maryland, 1–8. https://doi.org/10.1145/2811780.2811921

[38] W. D. Killgore. 1998. The Affect Grid: a moderately valid, nonspecific measure of pleasure and arousal. *Psychological Reports* 83, 2 (Oct. 1998), 639–642. https://doi.org/10.2466/pr0.1998.83.2.639

[39] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work (Lecture Notes in Computer Science)*, Andreas Holzinger (Ed.). Springer, Berlin, Heidelberg, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

[40] Richard S Lazarus. 2006. *Stress and emotion: A new synthesis*. Springer Publishing Company.

[41] Vincent Le Moign. 2017. Streamline Emoji, Free Icons from the Streamline Icons Pack. https://streamlineicons.com https://www.webalys.com/.

[42] Dominik J Leiner. 2019. SoSci Survey (Version 2.5.00-i1142) [Computer software]. https://www.soscisurvey.de/en/about Available at http://www.soscisurvey.com/.

[43] Boning Li and Akane Sano. 2020. Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (June 2020), 49:1–49:26. https://doi.org/10.1145/3397318

[44] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. MoodScope: building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services (MobiSys '13)*. Association for Computing Machinery, Taipei, Taiwan, 389–402. https://doi.org/10.1145/2462456.2464449

[45] P. F. Lovibond and S. H. Lovibond. 1995. The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy* 33, 3 (1995), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

[46] Douglas M McNair and Maurice Lorr. 1964. An analysis of mood in neurotics. *The Journal of Abnormal and Social Psychology* 69, 6 (1964), 620.

[47] Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology*. Cambridge (Mass.)[etc.]: MIT Press.

[48] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 75:1–75:21. https://doi.org/10.1145/3351233

[49] Greg Murray. 2007. Diurnal mood variation in depression: A signal of disturbed circadian function? *Journal of Affective Disorders* 102, 1 (Sept. 2007), 47–53. https://doi.org/10.1016/j.jad.2006.12.001

[50] John P. Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, Vancouver, BC, Canada, 725–734. https://doi.org/10.1145/1978942.1979047

[51] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (Sept. 2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[52] Lenore Sawyer Radloff. 1977. The Ces-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement* (June 1977). https://doi.org/10.1177/014662167700100306

[53] Verónica Rivera-Pelayo, Angela Fessl, Lars Müller, and Viktoria Pammer. 2017. Introducing Mood Self-Tracking at Work: Empirical Insights from Call Centers. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (Feb. 2017), 3:1–3:28. https://doi.org/10.1145/3014058

[54] Valentina Rossi and Gilles Pourtois. 2012. Transient state-dependent fluctuations in anxiety measured using STAI, POMS, PANAS or VAS: a comparative review. *Anxiety, Stress, and Coping* 25, 6 (Nov. 2012), 603–645. https://doi.org/10.1080/10615806.2011.582948

[55] James A. Russell. 1978. Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology* 36, 10 (1978), 1152–1168.

[56] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714

[57] James A Russell, Anna Weiss, and Gerald A Mendelsohn. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57, 3 (1989), 493.

[58] Kraig L. Schell and Frederick L. Oswald. 2013. Item grouping and item randomization in personality measurement. *Personality and Individual Differences* 55, 3 (July 2013), 317–321. https://doi.org/10.1016/j.paid.2013.03.008

[59] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[60] Stefan C. Schmukle, Boris Egloff, and Lawrence R. Burns. 2002. The relationship between positive and negative affect in the Positive and Negative Affect Schedule. *Journal of Research in Personality* 36, 5 (Oct. 2002), 463–475. https://doi.org/10.1016/S0092-6566(02)00007-7

[61] Sandra Servia-Rodríguez, Kiran K. Rachuri, Cecilia Mascolo, Peter J. Rentfrow, Neal Lathia, and Gillian M. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-being: A Large-scale Longitudinal Study. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 103–112. https://doi.org/10.1145/3038912.3052618

[62] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological Momentary Assessment. *Annual Review of Clinical Psychology* 4, 1 (2008), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[63] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2018. Mood modeling: accuracy depends on active logging and reflection. *Personal and Ubiquitous Computing* 22, 4 (Aug. 2018), 723–737. https://doi.org/10.1007/s00779-018-1123-8

[64] James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87, 2 (1980), 245–251.

[65] Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 715–724. https://doi.org/10.1145/3038912.3052676

[66] Boyuan Sun, Qiang Ma, Shanfeng Zhang, Kebin Liu, and Yunhao Liu. 2017. iSelf: Towards Cold-Start Emotion Labeling Using Transfer Learning with Smartphones. *ACM Transactions on Sensor Networks* 13, 4 (Sept. 2017), 30:1–30:22. https://doi.org/10.1145/3121049

[67] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017), 1–1. https://doi.org/10.1109/TAFFC.2017.2784832 Conference Name: IEEE Transactions on Affective Computing.

[68] Auke Tellegen, David Watson, and Lee Anna Clark. 1999. On the dimensional and hierarchical structure of affect. *Psychological Science* 10, 4 (1999), 297–303.

[69] Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242.

[70] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6 (Dec. 2017), 93:1–93:40. https://doi.org/10.1145/3123988

[71] Feion Villodas, Miguel T. Villodas, and Scott Roesch. 2011. Examining the Factor Structure of the Positive and Negative Affect Schedule (PANAS) in a Multiethnic Sample of Adolescents. *Measurement and Evaluation in Counseling and Development* 44, 4 (Oct. 2011), 193–203. https://doi.org/10.1177/0748175611414721

[72] Torben Wallbaum, Wilko Heuten, and Susanne Boll. 2016. Comparison of in-situ mood input methods on mobile devices. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia (MUM '16)*. Association for Computing Machinery, Rovaniemi, Finland, 123–127. https://doi.org/10.1145/3012709.3012724

[73] D. Watson. 1988. Intraindividual and interindividual analyses of positive and negative affect: their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology* 54, 6 (June 1988), 1020–1030. https://doi.org/10.1037//0022-3514.54.6.1020

[74] David Watson and Lee Anna Clark. 1999. *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form.* Iowa Research Online.

[75] David Watson, Lee A Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[76] David Watson and Auke Tellegen. 1985. Toward a consensual structure of mood. *Psychological bulletin* 98, 2 (1985), 219.

[77] David Watson and Jatin G. Vaidya. 2012. Mood Measurement: Current Status and Future Directions. In *Handbook of Psychology, Second Edition*. American Cancer Society, USA. https://doi.org/10.1002/9781118133880.hop202013

[78] Bert Weijters, Alain De Beuckelaer, and Hans Baumgartner. 2014. Discriminant Validity Where There Should Be None: Positioning Same-Scale Items in Separated Blocks of a Questionnaire. *Applied Psychological Measurement* 38, 6 (Sept. 2014), 450–463. https://doi.org/10.1177/0146621614531850 Publisher: SAGE Publications Inc.

[79] Xiao Zhang, Fuzhen Zhuang, Wenzhong Li, Haochao Ying, Hui Xiong, and Sanglu Lu. 2019. Inferring Mood Instability via Smartphone Sensing: A Multi-View Learning Approach. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, Nice, France, 1401–1409. https://doi.org/10.1145/3343031.3350957

[80] Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods* 12, 4 (2007), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

[81] Marvin Zuckerman. 1960. The development of an affect adjective check list for the measurement of anxiety. *Journal of Consulting Psychology* 24, 5 (1960), 457.