Mood Measurement on Smartphones
Which Measure, Which Design?

Torkamaan, Helma

DOI
10.1145/3580864

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies

Mood Measurement on Smartphones: Which Measure, Which Design?

HELMA TORKAMAAN, Delft University of Technology, Netherlands

Mood, often studied using smartphones, influences human perception, judgment, thought, and behavior. Mood measurements on smartphones face challenges concerning the selection of a proper mood measure and its transfer, or translation, into a digital application (app) that is user-engaging. Addressing these challenges, researchers sometimes end up developing a new interaction design and modifying the classic mood measure for an app. However, the extent to which such design alterations can impact user compliance, user experience, and the accuracy of mood measurements throughout a mood self-tracking study is unclear. In this paper, we explore and investigate how the selection of a mood measure (from two widely used measures) and its design alteration (from three options of classic, chatbot, and interactive designs) impact the (i) validity, (ii) user compliance, and (iii) user experience of mood measurement apps. For this purpose, we conducted a hybrid study with a mixed design in three parts. The first part suggests that a measure's validity can be susceptible to design modifications and introduces the concept of measure's resilience which can be essential when modifying the interaction design of a measurement tool. The second part discovers that both the type and design of the chosen measure can impact user compliance. This part also portrays a more complete picture of user compliance by demonstrating the use of several variables to investigate compliance. This investigation reveals that user compliance is not just about the response duration or length of a measurement tool. The final part finds that a measure or its design does not significantly influence the user experience for a well-designed app. In this part, we also discover which user experience criteria are more impactful for improving user compliance when designing mood tracking (or mood self-tracking) tools. Our results further suggest that, for a resilient measure, the interactive design is more likely to attract users and have higher user compliance and satisfaction as a whole. Ultimately, choosing a measure or design alternative would be a three-way trade-off between the measure's validity (or accuracy), user compliance, and user satisfaction, which researchers have to prioritize. A successful mood measurement with a smartphone needs to balance both concepts of app quality and assessment quality.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; **Usability testing**; **Touch screens**; **Empirical studies in HCI**; *User centered design*; **Smartphones**; *Mobile phones*; *Ubiquitous and mobile computing design and evaluation methods*; *Empirical studies in ubiquitous and mobile computing*.

Additional Key Words and Phrases: Mood Tracking; Interactive Design; Smartphone; User Compliance; Self-reports; Self-tracking; Experience Sampling Methods; Affect; Emotion; Chatbot;

ACM Reference Format:
Helma Torkamaan. 2023. Mood Measurement on Smartphones: Which Measure, Which Design?. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 29 (March 2023), 35 pages. https://doi.org/10.1145/3580864

1 INTRODUCTION

Affect — as a broad term for mood and emotion — has a fundamental role in the well-being, health, behavior, and cognition of an individual. Accordingly, measuring and tracking mood and emotion are attracting widespread interest in the community. Mood is a powerful and influential phenomenon to consider for various purposes, such as understanding, tracking, modeling, and predicting human behavior, decision, working performance,

Author's address: Helma Torkamaan, Delft University of Technology, Mekelweg 5, 2628 CD , Delft, Netherlands, h.torkamaan@acm.org.

2474-9567/2023/3-ART29 $15.00
https://doi.org/10.1145/3580864

well-being, or mental health. Researchers may track mood to help with health recovery or habit changes. They also monitor mood and emotion to design better products and even recommend personalized items to customers.

Affect research per se is a growing and complex field with over a century of scientific contributions. This has resulted in various definitions, theories, and measures of affect; and yet, one cannot find a unanimous definition for the basic concepts of this field, such as emotion [33] or mood. However, researchers seem to have more agreement on the distinction between mood and emotion [6] and in characterizing them. *Mood* — defined e.g., by [77, p.4] as "transient episodes of feeling or affect" — is generally longer lasting than emotion and has a span ranging anywhere from a few minutes to several hours or even days, while emotion is mostly anywhere between a few seconds to several minutes. Watson and Vaidya [81, p.351] have maintained that the center of attention in mood research has been the "subjective, phenomenological experience"; whereas emotion has been studied "as multi-modal psychophysiological systems" through the assessment of the following components: subjective, physiological, expressive, and behavioral (e.g., flight response). Table 1 summarizes some commonly discussed differences between mood and emotion in the literature [6, 20, 21, 28, 54, 57, 81]. Mood has relationships with, and may impact, several behavioral or psychological phenomena, including cognition and memory. Accordingly, it has been assessed and monitored in a wide range of studies targeting users' behaviors and abilities [66].

Table 1. Differences between mood and emotion derived from [6, 20, 21, 28, 54, 57, 81]. Other sources may discuss further differences or consider these characteristics differently.

Property	Emotion	Mood
Duration	Very short	Long but limited & not lifetime
Presence	Occasionally	Often
Rise	Quickly after stimulation	Slowly
Cause	Internal or external	Mostly internal
Intensity	High	Mild or low
Attention	Consume mind & attention	Background of the mind
Awareness of cause	High awareness	Possibility of awareness
About something	Yes	May not
Observable signals	Yes	No
Consequences	Mostly expressive and behavioral	Mostly cognitive

Assessing mood due to its subjective nature has been primarily based on traditional self-reported measurement instruments (or simply measures in this paper). In this paper, mood tracking (or mood self-tracking) is longitudinal repeated assessments of mood to detect and model mood changes or fluctuations. Mood tracking is typically conducted using experience sampling method (ESM). ESM or ecological momentary assessment (EMA) [59], is repeated, sometimes trigger-based (i.e. based on an event, schedule, or random triggering), sampling or assessment of users' states, feelings, behaviors, thoughts, etc. over time. For over a decade, smartphones have been the tool of choice for applying ESM [72] because of their convenience, and various platforms and applications (app(s)) have been developed for conducting ESM studies with them, e.g., [1, 13, 24, 69, 75]. Affect, in general, is the most commonly ESM-based tracked behavior using mobile devices [83].

Mood tracking using smartphones involves at least two types of decisions. Researchers first need to decide which measure to use. They should choose a measure based on several factors, such as their study's focus, the affect theory upon which the mood measure relies, its requirements including the target population, and its psychometric properties. These factors apply to every affect-related study; however, smartphone-based mood tracking additionally faces another challenge. Most proper mood measures are traditionally designed based on pen-and-paper questionnaires and therefore need to be transferred to an app. After selecting a measure, researchers need to think about the measure's design and how the user interaction with the measure and app

should look like. They may move away from traditional questionnaire-like interface designs and modify the user interface of, or interaction with, the self-reported mood measure, e.g., by presenting it in the form of a chatbot, or with a graphical input for various reasons, such as fulfilling the requirements of, or improving user engagement in, their study.

User compliance, which is defined as the share of completed questionnaires from the presented ones [72], is a driving factor in making decisions regarding the use of a measure and its design in any ESM study, including mood tracking studies. Engaging users throughout the study, preventing dropouts, and reducing missing, careless [19], or meaningless responses, are major challenges for longitudinal studies, including those with ESM components. The frequency and schedule of sampling, number of questions, study duration, and participants' compensation are mentioned in the literature as trade-offs impacting user compliance [73].

To increase user compliance, researchers are increasingly looking for ways to shorten measures [46, 55, 65, 68], and make them more attractive [60], mobile friendly [17, 39, 43, 71], and user-engaging. Changing the user interaction to a form of a chatbot [14, 27, 67], modifying the visual design [60, 71], and coming up with newly user-centered designed measures [17, 29], are examples of such efforts. Although the user experience with a mood tracking app might improve as a result of such efforts, the validity of modified or devised measures is mostly unexplored. The extent to which (if any) the design alteration of a measure influences user impression, experience, and compliance throughout an ESM-study in practice, is also unclear. As a consequence, even if the chosen mood measure is a traditional measure with excellent psychometric properties, it is unclear how the measure should be designed as an app in order to have a valid, accurate, reliable, usable, and engaging mood tracking solution.

In this paper, we address these gaps in the literature by looking into two types of mood measures based on the dimensional theories of affect and comparing their design alternatives to the classic (i.e. traditional) versions. We hypothesize that not only the measure but also its design and presentation can potentially influence the validity of the mood measure as well as the user experience and fatigue throughout the study. We compare the design alternatives and measures for their first impression, the accuracy of assessment, user compliance, and user experience. The design of the classic versions is as close as possible to the respective traditional pen-and-paper mood measure. For design alternatives, we chose a chatbot design and graphical user interface, which could be used in future ESM using conversational agents or interactive systems. We are specifically interested in investigating whether the type of a mood measure and its app design impact the self-reported mood validity (RQ1), user compliance over the time (RQ2), and the overall user experience (RQ3) of the ESM app and accordingly, the respective mood measure.

To answer our research questions, we conducted a study with three parts, each one addressing one of the research questions. The participants of this study used different types of mood measures and design alternatives. We provided a total of six mood tracking tools in a 2×3 study design: Two variations of mood measures and three design alternatives for each variation. These tools were developed and released using an Android app that tracks users' self-reported mood. This paper:

(1) Provides empirical evidence suggesting that mood measures can be different in their resilience toward interface modifications
(2) Presents and discusses comprehensive results showing how the selection of a mood measure and its design may influence validity, user compliance, and user experience
(3) Shows how user first impression of a measure (or tool) is different from their perception after use and compared to their actual behavior and engagement in practice
(4) Highlights that selecting and transferring a measure into a smartphone app is a challenging process that requires careful consideration of both assessment quality and app quality according to the measure's resilience

(5) Demonstrates a method for studying user compliance in ESM by proposing a set of variables and showing how the inclusion of these variables may lead to varying results

2 RELATED WORK AND BACKGROUND

2.1 Traditional Measures

Affect measures mainly capture either specific affective states or overall (the global or entire range of affective states) affect. This difference lies in underlying theories of affect. These theories, for the most part, have a discrete or dimensional view on affect. The discrete view on affect is a group of contributions in which researchers study specific affective states — deemed to be important — distinct from each other by their characteristics, evolutionary roles, and symptoms. For example, anger or depression are investigated and explained independently. Two example measures of this group are Profile of Mood States (POMS) [41] and Multiple Affect Adjective Check List (MAACL) [87]. These measures are restricted to specific states on which they focus; namely, POMS can only capture affective states of tension, depression, anger, vigor, fatigue, and confusion. They, accordingly, are unable to capture a complete picture of an individual's feelings or account for the interrelations of affective states, particularly in affective states with a similar valence, such as anger and sadness.

Unlike the discrete view on affect, the dimensional view does not focus only on specific affective states. It instead is about considering a multi-dimensional affective space, where each affective state is described by its relative position in this space, and, therefore, the whole spectrum of affective states can be characterized and monitored. Among various theories and models of this view, two are the most prominent (both two-dimensional) and are commonly accepted as the general dimensional models of affect. The first model identifies two dimensions of valence (sometimes called pleasure or pleasantness) and arousal (sometimes called energy or activations). We call it the pleasantness-energy model in this paper. Circumplex model of affect [53] is a well-established description of this model. The second model identifies two dimensions of positive affect (PA) and negative affect (NA) (sometimes called positive activation and negative activation), which resulted from a Varimax rotation of the dimensions in orthogonal factor analysis [79, 80]. In other words, PA-NA and pleasantness-energy dimensions are both portraying the same affective space.

There are a variety of measures for dimensions of pleasantness-energy and PA-NA, but some are used more often than others. Affect Grid [55], a single-item measure with a 2D–9×9 grid-like response format, is a notable measure of pleasantness-energy dimensions based on the Circumplex model of affect and is frequently used because of its shortness. Measures, such as Mehrabian and Russell [42]'s scale and Self-Assessment Manikin (SAM) [9], also can capture the pleasantness-energy dimensions. However, they additionally capture a third dimension, i.e. dominance, which is related to the historical development of the dimensional theories. SAM is brief and has been used when capturing an emotion or an individual's attitude toward an object, image, stimuli, etc. [9]. Positive and Negative Affect Schedule (PANAS) [79] with 20 items is the standard measure for capturing PA and NA dimensions. Later, Thompson [65] shortened this measure into a 10-item questionnaire called International-PANAS-Short Form (I-PANAS-SF). For mood tracking in general, or smartphone-based mood tracking in particular, the length of a measure due to the repetition of assessment and screen-size limitations is important. Other dimensional measures in the literature are either far lengthier, e.g., [23, 64, 78], or do not provide a body of psychometric data or evidence, e.g., [31, 34, 40], as extensive as the abovementioned measures. Accordingly, PANAS (or I-PANAS-SF) and Affect Grid have a widespread popularity as general dimensional measures of mood in the literature for assessing non-clinical populations.

Affect Grid and PANAS represent different dimensions of affect; however, this is not their only difference. Affect Grid is a short measure with only one item, and yet, it has a long instruction text with at least six different examples. In other words, using this measure requires thorough user training. In contrast, PANAS is a longer measure with 20 (or 10 for its short form, I-PANAS-SF) items, but it is self-explanatory and has a short instruction

text. Compared to any other affect measures, PANAS and its variants have an exceptional and diverse body of evidence, data, and research evaluating their psychometric properties and confirming their validity and reliability for capturing affect and mood, with various timing instructions, such as present moment, past day, week, etc., and for targeting various populations, in different languages. Affect Grid and PANAS are both used for mood tracking using smartphones, e.g., in [44, 51, 58, 61].

2.2 Modern Tracking Apps and Measures

Smartphone-based mood tracking apps might be based on recently devised affect measures, that are primarily developed for smartphones or designed to be user-engaging, and therefore can be used directly with smartphones without any modification. For example, Desmet et al. [17] presented an intuitive pictorial mood measure called Pick-A-Mood, or Hafiz et al. [29] designed a humor-based mood measure using smartphones. Several earlier works also tried to build or design smartphone-based mood measures, such as [39, 43]. The problem with most such mood measures is that they have not been evaluated for their psychometric properties or compared with traditional measures of mood. Conversely, Photographic Affect Meter (PAM) [46] is a smartphone-based affect measure that has been compared to traditional measures of affect, namely PANAS. It, however, only weakly correlates with NA value, despite containing negative items (i.e. pictures) and has not yet been compared to any measure of pleasantness-energy dimensions, which it has been designed based on originally, to the best of our knowledge. As a consequence, PAM does not seem to fully reflect the affective space. Limitations of a measure, in providing proper validation, question its effectiveness in measuring and monitoring general mood, particularly related to the neglected dimensions of mood.

Smartphone-based mood tracking apps could also use valid and reliable traditional measures of affect. However, the first challenge after choosing the proper measure is transferring it into an app, which may require design modifications to make the mood tracking tool usable and user-engaging for repeated assessments. Smartphones have functionalities far beyond pen-and-paper or desktop applications. As a result, the user interaction with the device and, correspondingly, the interaction with the measure can be unlike past studies. For instance, the visualization of the questions can be different since there is the possibility of using color, shape, and sound, as used in [35]. It is even possible to improve user experience by gamifying the questionnaires. Modifying the traditional measures could also be because of the requirements and context of the study, namely having a conversation-based interaction with chatbots [14, 26]. The changes in interactions or design may subsequently influence the measure function and its accuracy, validity, or reliability. Therefore, unless the modified measure is validated or evaluated against its unmodified (or classic) form, the captured mood values and, accordingly, any developed models based on it cannot be relied upon or reproduced across various studies.

However, investigating the validity of a mood measure after its modification is not prevalent. Dubad et al. [18] systematic review of mobile mood-monitoring applications in young people found that only 36% of the studies in their review have reported the validity or reliability of the apps. This figure could be far worse in non-clinical studies or those that are published in technical venues, such as ACM and IEEE, where often the psychometric properties of the devised or modified mood tracking tools are not discussed at all. Studies that follow a measure or its user interface alteration with a validation, such as [68] or the effort of Sonderegger et al. [60] in improving the SAM manikins and that of Ugur et al. [71] in looking into a grid-like 2D self-report similar to Affect Grid and its visual design alternatives as two independent linear scales, are scarce in these communities. One can even find studies, e.g., [3, 14, 38, 62, 63], that do not use the typically discussed concepts of mood in the affect literature, and instead use a self-defined arbitrary measure of mood with items, such as good, bad, fine, neutral, etc. Such mood assessments are inexplicable in terms of affect models, theories, and their respective measures. This highlights the need to investigate mood tracking apps, compare them, and explore their characteristics and validity.

In short, although the affect literature is rich in comparing measures concerning their underlying theories, dimensions that they use, and what they essentially measure, e.g., [52], there is still a need for investigating them as tracking apps, particularly for their user experience and validity (if modified). Wallbaum et al. [74], therefore, compared four mood measures — supposedly all based on the Circumplex model of affect [53] but mostly modified or PAM-related — in self-reported intuitiveness, inconvenience, speed of input, everyday use, expressiveness, and overall suitability with a small user study. Nevertheless, neither user compliance nor the validity of those measures was investigated.

2.3 User Compliance in Self-tracking Apps

Tracking of affect, in general, has been reported as the most frequently, primary, and secondary, targeted behavior in ESM-based studies using mobile devices [83]. There is also an increasing body of literature on ESM-based studies [72, 82, 83] as well as those with insights and methodological contributions investigating impactful factors on user compliance, e.g., [7, 8, 13, 47, 50], and testing various ESM techniques concerning user experience and compliance, e.g., using unlock-journaling [86], microinteractions [48, 85], widgets [15], and on-body devices [2]. However, these studies are often associated with a basic consideration of user compliance, and other possible variables of compliance are mostly unexplored.

The definition and reporting of compliance in ESM studies are also inconsistent. User compliance is reported only in about 40.9% [72] to 57.7% [83] of ESM-based studies. When reported, sometimes different definitions of compliance are used [82]. For example, often, the classic definition of compliance is considered as the number of completed questionnaires in an ESM study [72]. Sometimes, it is considered as the share of completed sampling events among those that the users indeed have received. Intille et al. [32] differentiated between the two, calling the latter completion rate. Nonetheless, these definitions alone may not be sufficient to provide a complete picture of user behavior concerning overall compliance in a longitudinal mobile-based study. For instance, they do not count for the meaningless responses or responses that have been submitted too quickly. Response duration, quality of responses, dropouts, and endurance could be important metrics.

Various factors besides the measure, its length, and app design can impact user compliance. The study design, frequency of sampling, study duration, compensation, and other study tasks are a few examples of these factors. For the abovementioned reason, there is no clear baseline to foresee or compare user compliance of measures or their design alterations in general. It is also difficult to find a comparison between the newly developed tools and classic measures in user compliance, or even accuracy or user experience, of the measure, possibly due to the lack of a baseline for the classic tools. As a consequence, the extent to which the length and design alterations can impact user compliance is still unclear. In short, this is partly related to the lack of adequate comparisons with empirical evidence between measures or designs.

Altogether, to the best of our knowledge, there has been no comprehensive comparison between PA-NA and Pleasantness-energy dimensional measures, such as PANAS and Affect Grid — two of the most widely used measures of affect — and their design alternatives using smartphone apps. There has been little discussion about how modification of the user interface or interaction design may impact the quality of assessment, validity, user compliance, and user-perceived usability of a mood tracking app in general. Despite usability or user experience improvement claims, there is currently no baseline with which one can compare a newly designed or modified measure, either. The user experience of a measure has been assessed mostly after one-time use but rarely after longitudinal and repeated use. Therefore, it is unclear if users' first impressions would differ from their experience after repeated use and actual compliance behavior. This study seeks to address these gaps by investigating how a chosen mood measure or its design alternatives influences the validity, user compliance, and user experience of a mood tracking app.

3 OVERALL STUDY DESIGN

The study design has three parts. Each part investigates one of the research questions: *Part 1*: Pre-study addressing RQ1 (i.e. concurrent validity and comparability of the measures and design alternatives) with a one-time within-subjects assessment; *part 2*: The main study evaluating RQ2 (i.e. comparing user compliance variables), using ESM of two weeks with a between-subjects study design; and *part 3*: Post-study investigating RQ3 (i.e. comparing user experience and usability of each measure and design alternative) with a between-subjects study design. For our investigations, we developed and released an Android app that tracks users' self-reported mood. For this study, the app contains a total of six mood measurement tools based on two types of measures and three design alternatives for each type (2×3 study design) and would be open to researchers for future use upon publications[1]. By installing the app, every user could participate in the study depicted in figure 1. In this section, we explain the six mood measurement tools (i.e. two measures and their three design alternatives). The details of the study for each research question, their results, and discussions follow in the next sections (sections 4, 5, and 6). We then further discuss the study results and implications altogether in section 7.

Fig. 1. User journey of participating in the study.

3.1 Mood Measures

We considered two measures as representatives of the most prominent dimensional approaches, i.e. PA-NA and pleasantness-energy: I-PANAS-SF scale [65] based on PANAS [79] denoted as **A** in this paper, and Affect Grid [55] denoted as **B**. We chose these measures because of several reasons and requirements. One reason is that these measures are among the most popular tools being used in the domain by both psychologists and computer scientists for the general assessment of mood. Another reason is that the focus of this study was to improve the longitudinal assessment of mood as an overall feeling, and not as a set of discrete emotional states. As a result, dimensional approaches were suitable here for capturing an individual's mood. Finally, for longitudinal repeated measurements, the length of the measure (which depending on the measure can vary between one to 132 items) could be of particular importance and limits the pool of possible choices of general-purpose mood measures for non-clinical populations. Measure A and B are both relatively short compared to other available measures, which makes them suitable for repeated assessments.

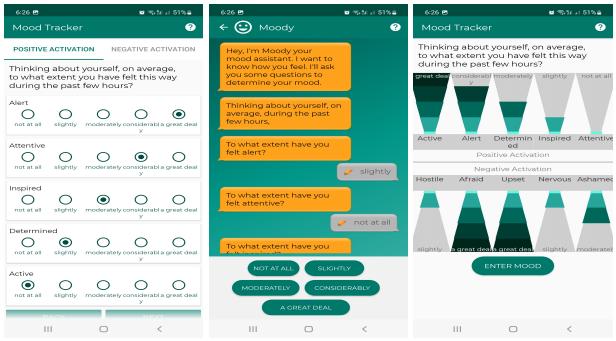
3.2 Mood Measure Design Alternatives

We studied three design alternatives for each of the considered mood measures: **(A)** I-PANAS-SF and **(B)** Affect Grid that utilized various interaction styles: **(1)** classic; **(2)** chatbot; and **(3)** interactive touch-based. The two

[1]The application is available for non-profit academic research purposes and can be requested here: https://torkamaan.de/#download. Several measurement tools were developed for this app as part of a bigger project. Some of these tools were devised affect measures — intended for answering different research questions — that are outside the scope of this paper. As an example, concise and adaptive measurement instruments were studied using this research app [68].

measures of A and B and their design alternative, *1, 2,* and *3*, together shaped the six mood tracking tools of this paper: $A_1, A_2, , A_3, B_1, B_2$, and B_3[2]. Figure 2 shows screenshots of the three interaction styles of this study.

(a) Measure group A: PA-NA-based tools



(b) Measure group B: Pleasantness-energy-based tools

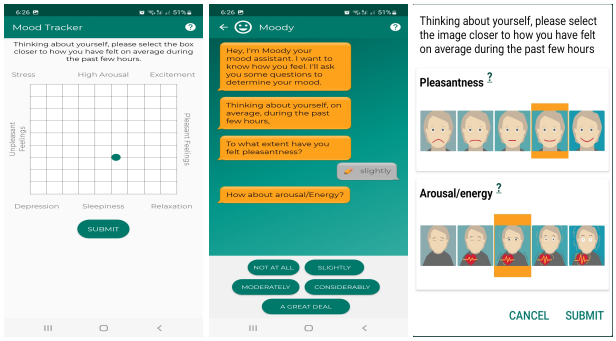


Fig. 2. The screenshots of the six mood tracking tools (2×3 study design). On the left, A_1 and B_1: classic. In the center, A_2 and B_2: chatbot, and On the right, A_3 and B_3: interactive.

For measure group A, based on I-PANAS-SF, the first interaction style or classic (A_1) tried to keep the measure as close as possible to the traditional pen-and-paper based questionnaire. For example, it had radio buttons and Likert-type scales. The second interaction style (A_2) provided a chatbot interface. The final interaction style (A_3) used visualization and touch movements to provide a graphical interactive experience.

For measure group B, Affect Grid based on pleasantness-energy dimensions, we considered Affect Grid's original form as the classic version (B_1). Affect Grid seems to function similarly to single-item pleasantness and single-item energy measures according to very strong correlations reported by Russell et al. [55]. We accordingly designed the chatbot version (B_2) by asking about pleasantness and energy with single-item questions similar to [55]. For the interactive version (B_3), we represented the pleasantness and energy via manikin faces inspired by SAM [9] and [60], yet still reflecting single-item questions.

[2]The design alternatives were developed following iterative design cycles to assure their usability. Several options were iteratively prototyped (different levels of fidelity), refined, and evaluated with several small groups of test users independent of this study (total $n = 17$) using co-creation, observations, interviews, and usability questionnaires, which their details are outside the scope of this paper. Ultimately, one option per category of the design alternative was chosen based on the users' feedback as the final design.

3.3 Participants Recruitment and the Application Release

We announced the study[3] using a press release that described a two-week-long study on mood using the app. Each participant could join the study by following the provided link and installing the app from the Google Play Store, completing an onboarding journey, and agreeing with the participation consent and conditions, such as the minimum age. Participants were informed of the study description and their GDPR rights through the onboarding process. They did not receive any financial compensation for their participation. Our study was specified for participants older than 18 years old, and accordingly, all invalid age-related entries were discarded from the collected dataset later through a data cleaning process.

4 PART 1: RQ1-VALIDITY

4.1 Method: RQ1-validity

The goal here was to compare the six mood tracking tools in terms of the accuracy of the submitted values and the users' first impression of them in a one-time assessment. Formally, we hypothesized that the design alternatives of a measure have no impact on the resulting values of mood. Since the mood experience is subjective, we could compare the submitted values of mood between various measures or design alternatives only with a within-subjects study design.

After joining the study, users faced the pre-study (i.e. first part; within-subjects) with two optional tasks. The first task asked users to enter their mood for the past few hours, using all measures and their design alternatives concurrently, one following another in a random order (figure 2)[4]. Using the instruction of *past few hours* consistently, and assuming that mood last at least for a few minutes to several hours or even days as described in section 1, we assured that a user's mood was recorded with all six tools (2 measures × 3 design alternatives) in this task. Therefore, the design alternatives of each measure should ideally provide the same values for each user, assuming that design alteration has no impact on the measure's validity.

The second task asked users to specify their preferences of the measurement tools by first stating which one they liked the most. It next required users to choose the measurement tool they would have been the most satisfied with if they were supposed to use it repeatedly; and finally, users also selected the tool they disliked the most. In total, 452 users (female: 319, male: 121, other options: 12; with an average age of 32.42 (SD = 11.31) years) completed the pre-study in its entirety.

4.2 Results: RQ1-validity

4.2.1 Measure Group A (I-PANAS-SF based on PA and NA dimensions). Table 2 shows the description of mood values and the Spearman correlation among the two measures and their design alternatives. For group A as a representative of PA-NA dimensions, there were strong intragroup correlations. Both chatbot (A_2) and interactive (A_3) variations had very strong significant correlations with the classic design (A_1) for PA and NA variables. A_2 and A_3 also had strong significant correlations with each other. Furthermore, design alternatives of group A (i.e. A_1, A_2, and A_3) had correlations with almost equal strength with B_1, as well as with B_2 and B_3 (i.e. design alternative of group B). The correlation between PA and NA variables for all design alternatives of group A were also similar and within the range, but on the lower side, of the reported values in the literature [65, 79]. Internal consistency reliabilities (Cronbach's coefficient α) were all acceptably high, taking design alternatives together (PA: α = .951, NA: α = .927), pairwise, and separately (α > .851). All things considered, the design variations of this group could be used safely instead of each other.

[3]This study was assessed and approved by the institutional ethics committee at University of Duisburg-Essen.

[4]In addition to the described measurement tools in this paper (section 3.2), the users entered their mood using three more tools in this task. These additional tools were all devised affect measures, and their discussion is beyond the scope of this paper.

Table 2. The mean, standard deviation, median, and Spearman correlations of the mood values among measure groups A and B and their design alternatives.

Measure			(A) I-PANAS-SF						(B) Affect Grid					
Variable			PA			NA			PL			EN		
Design			1	2	3	1	2	3	1	2	3	1	2	3
	M		2.46	2.44	2.33	2.01	2.01	1.97	2.69	2.20	2.95	2.84	2.40	2.55
	SD		0.89	0.87	0.90	0.82	0.85	0.83	1.16	1.04	1.10	1.12	1.05	1.01
	Mdn		2.4	2.2	2.2	1.8	1.8	1.8	2.5	2	3	3	2	2
A	PA	1												
		2	0.88											
		3	0.85	0.86										
	NA	1	-0.26	-0.24	-0.23									
		2	-0.30	-0.27	-0.25	0.92								
		3	-0.30	-0.27	-0.21	0.90	0.90							
B	PL	1	0.47	0.46	0.46	-0.61	-0.64	-0.62						
		2	0.45	0.44	0.43	-0.32	-0.30	-0.35	0.43					
		3	0.55	0.51	0.53	-0.57	-0.59	-0.58	0.65	0.50				
	EN	1	0.37	0.36	0.37	0.24	0.22	0.22	-0.06	0.15	0.08			
		2	0.62	0.63	0.61	-0.20	-0.23	-0.23	0.40	0.47	0.41	0.37		
		3	0.48	0.50	0.48	-0.01	-0.02	-0.01	0.18	0.23	0.30	0.43	0.56	

Note: All values in black in the correlation table are significant; $df = 450; p < .001$. A: I-PANAS-SF representing PA-NA and B: Affect grid representing pleasantness-energy dimensions; 1: classic, 2: chatbot, and 3: interactive; PL: pleasantness, En: energy, PA: positive affect, and NA: negative affect.

4.2.2 Measure Group B (Affect Grid Based on Pleasantness and Energy Dimensions). Nevertheless, the same conclusion does not apply to group B, which captures the values of pleasantness (PL) and energy (EN) variables. The design alternatives of this group had only moderate yet significant correlations with each other. They also had such moderate correlations with group A (A_1, A_2, and A_3) although they capture different variables (as discussed earlier in section 2, dimensions of PA-NA appear to be a 45° rotation of pleasantness-energy dimensions in the affective space). In addition, comparing the correlations between the design alternatives of group B revealed a difference between dimensions of pleasantness and energy and, subsequently, their respective variables, i.e. PL and EN. For PL, the correlation between the classic (B_1) and interactive designs (B_3) was higher than the correlation between other design alternative pairs; whereas, for EN, the correlation between the chatbot (B_2) and interactive designs (B_3) was higher. In other words, design alteration in group B did not necessarily lead to consistent mood values.

4.2.3 Within Group Differences. In group A, a Friedman's test showed a significant difference among the design alternatives (A_1, A_2, and A_3) for PA variables ($\chi^2_F(2) = 36.97, p < .001; Kendall's - W = .041$) and for NA variables ($NA - A : \chi^2_F(2) = 13.136, p = .0014; Kendall's - W = .015$). The near-zero values of Kendall's-W for both variables, together with high correlation coefficients among A_1, A_2, and A_3, indicated a very small effect size for the observed difference between the design alternatives. This could be linked to the relatively large sample size for the within-subjects study design in this part. To examine where the difference actually lies, we applied post hoc analysis using Wilcoxon signed-rank tests with a Bonferroni correction resulting in a significance level set at $p < .003$. Interestingly, there was no significant difference between the classic and chatbot designs nor between the interactive and chatbot designs for the NA variable. However, there

were statistically significant reductions in PA values of the interactive design compared to both the classic and chatbot design alternatives, as well as in NA of the interactive design compared to the classic design ($PA_{(A_1,A_3)} : Z = -5.04; PA_{(A_2,A_3)} : Z = -6.066; NA_{(A_1,A_3)} : Z = -3.197; p < .001$).

Similarly, group B showed a significant difference for both PL ($\chi^2_F(2) = 174.655, p < .001; Kendall's - W = .193$) and EN ($\chi^2_F(2) = 46.148, p < .001; Kendall's - W = .051$) variables. Post-hoc analysis of group B revealed a significant difference between all designs ($-12.328 < Z_{PL} < -5.678, -7.359 < Z_{EN} < -3.341, p < .001$). The slightly higher value of Kendall's-W for PL suggests more agreement on the design's orders and accordingly indicates a bigger difference between the design alternatives in measure group B.

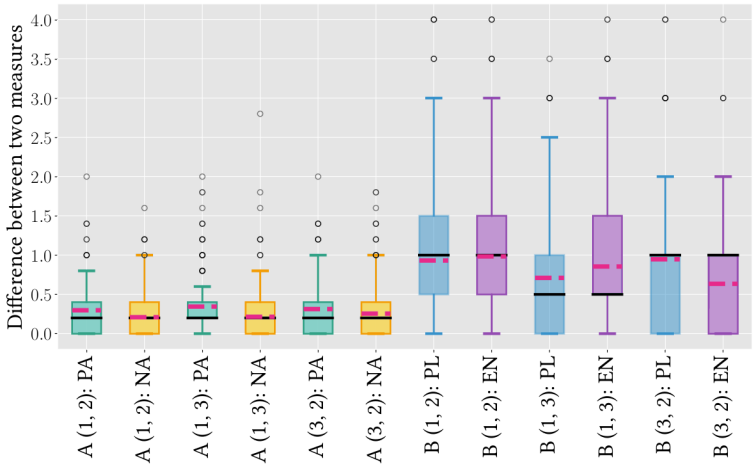


Fig. 3. Distance between the design alternatives. A: I-PANAS-SF representing PA-NA dimensions, B: Affect grid representing pleasantness-energy dimensions, 1: classic, 2: chatbot, 3: interactive, PA: positive activation, NA: negative activation, PL: pleasantness, EN: energy. The pink dashed line shows the mean, the black line is the median

4.2.4 Design Alternatives Agreement. To further investigate the difference between the design alternatives in each measure group as an indicator of the mood value inconsistency or error, we calculated the total Manhattan distance. In group A, the error was considerably low (the maximum mean distance of the design alternatives: $M_{A_3,A_2} = .57, SD_{A_3,A_2} = .46$) and below one unit of measurement. Group B generally had a higher inconsistency of the captured mood values (for its design alternatives, the maximum mean distance: $M_{B_1,B_2} = 1.92, SD_{B_1,B_2} = 1.15$ and the minimum mean distance: $M_{B_1,B_3} = 1.56, SD_{B_1,B_3} = 1.1$). The smaller the resulting inconsistency value (i.e. Manhattan distance), the closer and more similar the two measures are.

We additionally considered each mood variable independent of its pair value (e.g., comparing PA and NA) in a more detailed analysis. When comparing the error of PA and NA (or PL and EN for group B) for each pair of design alternatives, the distributions of error appeared to be the most similar for chatbot and interactive designs in group A, and for classic and chatbot design in group B. Figure 3 shows the mean, median, and box plot of this distance (mood value inconsistency) for each variable. Regardless of the mood variable, group B had a higher error compared to group A, and its errors had wider distributions and were more dispersed. The high error of group B compared to group A was not related to a particular variable and consistently applied to all variables.

4.2.5 Mapping Measure Groups. Group A (A_1, A_2, and A_3) and B (B_1, B_2, and B_3) basically were measuring varying variables respective to their underlying dimensional models. As also mentioned by Russell et al. [55],

one would expect the PA and NA variables of group A to have consistent correlations with PL and EN variables of group B — i.e. the correlations of PA and NA with EN are equal, and the correlation of PA and NA with PL are equal in absolute value, but the NA correlation with pleasantness is negative in sign. However, regarding PL, the strength of its correlation with NA appeared to be stronger than with PA for B_1 (classic design) and lower for B_2 (chatbot design), though it was not the case for B_3 (interactive design), where PL had almost a perfect relationship with PA and NA variables. Concerning EN, — similar to [55]— we could not find an accurate mapping of variables and equal correlations between PANAS and Affect Grid.

In addition, since our classic design of group B — i.e. Affect Grid — also had labels of excitement, depression, stress, and relaxation (see figure 2.(b)) as its original design, we also tried transferring its variables into PA and NA scores with a similar approach to [46]. The resulting PA and NA values for all tools ($A_1, A_2, A_3, B_1, B_2, B_3$) were significantly different from each other (PA: $\chi^2_F(5) = 1492.594, Kendall's - W = .660$ and NA: $\chi^2_F(5) = 914.889, Kendall's - W = .405; P < .001$). A pairwise comparison of all design conditions from both measure groups A and B, excluding within-measure group pairs, showed a significant difference. As indicated by the correlation values and unsymmetrical behavior of the outcome variables of Affect Grid, i.e. PL and EN, transferring these variables to PA and NA scores would not lead to an accurate estimation and, therefore, should be avoided.

4.2.6 Second Task: User First Impression. Users' measures and design alternatives of choice were diverse when after one-time use, they were asked to choose from the mood tracking tools in the second task. Choosing a measure that they liked the most, they mostly (21.02%; $n = 95$) chose A_3, i.e. the interactive design of measure group A, followed equally (14.38%; $n = 65$) by A_1, i.e. the classic version of group A and B_3, i.e. the interactive form of group B. About 46%; $n = 208$ of users chose the same measures when asked about repeated use in the second question, and the phrase of *using the measure repeatedly* did not change the overall user choices that much. The main difference, however, was that more users selected A_1 (16.15%; $n = 73$) compared to B_3 (15.71%; $n = 71$) this time. More users (15.71%; $n = 71$) also chose B_1, i.e. the classic design of group B, Affect Grid, compared to the first question (12.61%; $n = 57$), as well. At the same time, the third and last question revealed that (30.09%; $n = 136$) of users disliked B_1 the most, followed by A_3 (13.05%; $n = 59$). The chatbot design of both groups A and B (A_2 and B_2) was both liked and disliked the least compared to other design alternatives, and therefore was the least attractive or stimulating design alternative.

4.3 Discussion: RQ1-validity

One of the main goals of the first part was to investigate the resilience of mood measures to design changes and the impact of such design changes on the resulting values of mood. The results both negated and supported the hypothesis and revealed the relationship of the measures and their design alternatives with each other.

Measure group A — a measure of PA-NA dimensions — overall seemed very stable and resilient toward interaction and design changes. The strong concurrent validity of the chatbot and interactive designs of group A suggested that they were accurate and could lead to similar results compared to the classic version. The chatbot design appeared to be more similar to its classic version compared to the interactive design, although only slightly. The error of measurement within this group was low and probably would not considerably influence the outcome of a study; however, one should consider that the interactive design tends to attenuate the mood values to a small degree compared to its classic version. From another perspective, users overall liked the interactive design more than the chatbot in a one-time assessment.

Measure group B — a measure of pleasantness-energy dimensions — contrary to group A, did not appear to be flexible to any design changes. Design modification in this group resulted in inaccurate outcomes, unlike measure group A. Overall, the design alternatives of group B seemed to have lower accuracy. The associations between the resulting PL and EN variables from the classic design of group B with other design variations of this group were inconclusive. Similar associations, even with stronger strength, were seen with PA and NA variables

of measure group A. PL generally appeared to be more robust than EN in Affect Grid, as the strength of the association between the classic design and its alternatives in group B for the PL variable was consistently higher than that of the EN variable.

There can be several reasons for the surprisingly high observed errors of group B, which partly conflicts with earlier research carried out in this area using pen-and-paper-based measures [55]. According to Russell et al. [55], one expects to observe a very strong correlation between single-item pleasantness-energy and Affect Grid; yet, no such association was found either for PL or EN variables. One may argue that group B is more suitable for capturing emotion rather than mood, as the single-item pleasantness-energy questions were tested with a set of emotional words by Russell et al. [55], thus making it weaker in reflecting dimensions of an individual's experienced feelings. Another reason could be related to the complexity of Affect Grid. The required user training for this measure, and the fact that users generally disliked this measure the most among other choices (task 2 - user first impression 4.2.6), suggest the possibility of having inaccurate user inputs. Since this measure is not that self-explanatory, users may have more difficulty with it even after training and could, therefore, record inaccurate or meaningless responses at first. Finally, one should also consider that this observation could be related to the characteristics and limitations of Affect Grid and its inflexibility for any modifications. Taken together, this measure should be carefully used and preferably even using the original pen-and-paper settings where possible. When modified, this measure should be thoroughly tested before use to ensure the quality of the assessment.

The remarkable result to emerge from this investigation is that the design of a measure can influence the resulting values depending on the resilience or robustness of a measure. For a measure like I-PANAS-SF that has self-explanatory items and is easy to understand, interactive or chatbot-based design variations did not change the outcome significantly. In contrast, Affect Grid is a complex measure, but its design alterations led to varying outcome variables. Transferring the values of this measure, or its design alternatives, to PA and NA dimensions would also lead to inaccurate, if not meaningless, estimations. Section 6 looks into the complexity and user experience of these mood tracking tools further.

5 PART 2: RQ2-USER COMPLIANCE

5.1 Method: RQ2-user Compliance

The second part focused on mood tracking over time with repeated measurements, in which user compliance plays an important role. The goal was to determine whether the type and design of a mood measure influence user compliance in a longitudinal assessment of 14 days. We used the six mood tracking tools (figure 2) described earlier (section 3.1) with a between-subjects study design.

5.1.1 Procedure. After joining the study and completing the pre-study (or part 1 of the study described in section 4), participants were randomly assigned to one of the six user groups, each using only one of the mood tracking tools ($A_1, A_2, A_3, B_1, B_2, B_3$) for two weeks. The app promoted sampling events at least two times per day. Although participants were instructed to use the app and try answering its questions for at least two weeks, they could continue using it as long as they wanted. Consequently, around 73 users kept actively using the app after over a year.

5.1.2 The App Description. To compare user groups accurately, the app had similar functionalities across user groups. According to a pre-defined schedule (sampling event), the app showed a mood measurement pop-up on the screen together with an ongoing notification until the measurement was either finished or dismissed. The sampling events were presented at least twice daily (section 5.1.3). Each sampling event was tagged as either *completed*, *dismissed* (i.e. if a user did not engage with the sampling event and dismissed it immediately), or

invalid (i.e. when the user started answering but did not finish and exited the sampling event in the middle)[5]. The app also provided feedback to users on the main screen of the app consistently for all user groups. This feedback summarized the users' overall mood of the day and the mood of weekdays, as well as a diagram of the users' mood history in a separate window. By showing the reports to the users as proposed by Hsieh et al. [30], we intended to engage them and minimize the dropouts throughout the study.

5.1.3 Frequency and Occurrence. We limited the number of sampling events to two times per day within the pre-defined timeframes to prevent user fatigue over time and capture mood fluctuations. Through the app settings, users could optionally schedule a third sampling event as well. The first sampling in the day was in the mornings (9:00–12:59 h), in which the average mood, particularly positive affects, might be at its highest level [16, 45, 84]. We had another one after lunchtime (14:00–15:59 h) to detect a possible drop of mood and, finally, an optional sampling in the evenings (19:00–21:59 h). Users were able to modify the exact default time of the sampling event only within the pre-defined timeframes via the app settings. They could also enter their mood manually and on-demand through a button designed on the main screen.

5.1.4 Variable Description. Compliance in the literature [25, 32, 72, 82] has been mainly considered as the number of valid and successfully submitted responses against the total number of samplings in an ESM. However, this working definition per se may be insufficient in finding the differences between the designs; for instance, it does not count for the meaningless responses or responses that have been submitted too quickly. Therefore, we extended this working definition of compliance by considering additional variables: *basic compliance*; *weighted compliance*; *completion rate*; *dropouts*; *endurance*; *motivational response*; and *total time spent* on a tracking tool.

The *basic compliance* is the count of completed (only from a scheduled sampling event) against the total sampling events required. In contrast, the *weighted compliance* does not just take the number of completed events and further considers response duration as a denominator factor. Accordingly, it taxes entries that have been recorded too quickly than reasonably possible ever to record one's mood using any measurement tool. For i-th sampling event ($i = 1, ..., n$) with response-duration of x_{ij}, of each user j, the weight of the sampling event w_{ij} is calculated using the absolute deviation of x_{ij} from the users' response-duration median ($\widetilde{X_j}$). Using the user's response-duration median instead of the user group's response-duration median in the weight function can count for the individual (or subjective) differences in response duration in addition to the measurement tool (assigned measure or design) differences. This weighting function only penalties responses that are submitted too fast, subsequently, $w_{ij} = (\left|x_{ij} - \widetilde{X_j}\right|)^{-1}$ where $x_{ij} < \widetilde{X_j} - 1$ and otherwise, it is equal to 1.

For example, consider Alice (54 Y), who has been assigned to use B_2. The median response duration of all her submitted samplings is 7 seconds ($M = 9.79, SD = 8.30, Min = 3, Max = 35$). Take an entry that was submitted in 4 seconds. In basic compliance, every completed entry, including this one, would be counted as 1, resulting in a total of 20. The weighting function in the calculation of weighted compliance taxes the responses that are submitted quicker than 6 seconds, depending on how quick they were[6]. As a result, this entry would be counted as 0.34. Summing all weighted entries and rounding them to the closest integer resulted in the value of 16 as the weighted compliance count for Alice.

We defined the rest of the variables as the following. The *completion rate* is the number of completed samplings (only scheduled) against the total sampling events a user has received. This value could be different from the

[5]At the end of each twice-daily mood sampling event, a self-reported single-item stress measure, similar to [76], was presented to all users similarly regardless of their assigned user group. Once per week, on Sundays at 18:00, an optional mood measurement — asking for the user's mood of the past few days using a hybrid measure — was also presented similarly to all users. The resulting values of the once-per-week mood assessment and perceived stress were recorded and tagged independent of the twice-daily sampling events. These additional assessments were intended and used to investigate other research questions which are not discussed in this paper.
[6]Values were rounded.

compliance variable. For example, if users' phones are off, they would receive no event; therefore, it would not impact the *completion rate*, unlike the *compliance rate*. *Dropout*, as another variable, shows how many full days up until two weeks (the study period) a user had participated before they dropped out. In contrast, the *endurance* variable counts the days of participation even after the two weeks. We additionally counted manual user entries as the representation of the *user's motivation*. This variable is independent of other variables concentrating only on the scheduled events. Manual entries occur when users open the app by their own choices and choose from the main screen to record their mood states. Finally, we also calculated the *total time* variable to compare the total time users spent with the measurement tools.

5.1.5 Data Preparation. To prepare the collected data for analysis, we were strict with a set of pre-defined conditions to have a homogeneous range of participation for all user groups and to exclude participants who did not engage with the study in the first place. We excluded all participants who dropped out within the first day of their participation in our assessments. Nevertheless, we still looked into the excluded users ($n = 165$) separately and found no significant evidence of an impact or difference between user groups for these users. These users were still considered for visualizing the dropout variable. We also excluded mood entries submitted after the 14th consecutive day of participation (after the installation date) to keep a comparable participation period between the users for all variables except the *endurance* variable. This variable, by definition, counts for days of participation beyond the study period. We subsequently calculated all variables for the included cases. From a total of 540 users assigned to six user groups, 375 (female: 259, male: 109, other options: 7) remained after data cleaning. The participants' age ranged from 18 to 75, with an average of $M = 33.47$ ($SD = 11.86$) years old.

5.2 Results: RQ2-user Compliance

Table 3. The description of the various compliance-related variables across the measures and design alternatives.

Groups		n	response duration			# completed	Compliance count			Weighted compliance			Completion rate			Drop out			Endurance			Motivational			Total time		
			M	SD	Mdn	events	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
A	1	55	942	2213	53	1356	9.7	11.4	5	5.7	6.4	3	0.4	0.3	0.3	8.9	5.1	9	32.6	44.9	12	1.9	4	1	4.6	5	3
	2	64	728	2008	40	1859	10.8	11.3	6	6.8	6.7	4.5	0.4	0.3	0.3	10.7	4.7	14	30.6	35.9	15	2.1	2.4	2	3.2	4	2
	3	62	767	2023	32	1880	12.5	13.6	6	7.7	8.2	4	0.4	0.4	0.4	10.5	4.8	14	48.2	58.4	24.5	2.8	3.8	1	4.1	4.5	3
B	1	68	790	1815	20	2317	18	14.3	18	11.4	8.8	12.5	0.5	0.3	0.5	10.7	4.7	14	58	64.5	29	3.4	4.7	2	5	5.2	4
	2	66	949	2331	19	2088	18.8	13.8	23.5	12	8.9	14	0.6	0.3	0.7	10.9	4.9	14	44.8	52.9	20.5	2.4	2.8	2	5.9	6.6	4
	3	60	902	2101	14	1749	15.9	12.4	14.5	10.7	8.3	9.5	0.6	0.3	0.6	10	5.3	14	48.1	56.8	30.5	3.1	3.7	2	5.7	5.6	4

Note: response duration values are in seconds. The compliance count (or basic compliance) is the count of completed entries from scheduled sampling events. Weighted compliance shows the count considering the response duration. Completion rate is concerned with events that the users have actually received. Drop out represents the total days of participation within the two weeks of the study. Endurance is the total days of participation even after two weeks. Motivational represents the motivational response-count entered by users manually. Total time shows the overall time-spent with the tool in hours. Measure group A: the measure of PA-NA dimensions, i.e. I-PANAS-SF, and measure group B: the measure of pleasantness-energy dimensions, i.e. Affect Grid; 1: classic, 2: chatbot, and 3: interactive design alternatives.

In total, 11,249 events were completed in this part, while the classic design of Affect Grid (B_1) had the most ($n = 2,317$), and the classic design of I-PANAS-SF (A_1) had the least ($n = 1,356$) number of entries. Table 3 provides a summary of all compliance variables of this experiment. Measure group B (B_1, B_2, and B_3) seems to exhibit higher compliance as a whole compared to measure group A (A_1, A_2, and A_3) (1: classic; 2: chatbot; 3: interactive designs).

5.2.1 Compliance Basic and Weighted. Variables of *basic* and *weighted* compliance both captured compliance counts, and the analysis procedure for them was similar. These count-based variables presented repeated assessments for each user and had skewed distributions. We first ran a preliminary standard Poisson model (using the

log-likelihood function) for the compliance counts based on the user groups and other covariates and determined the dispersion status of the model. We found that none of the included covariates, such as age or response duration, had significant effects in the model. As a consequence, we fitted the model simply based on the user groups. The log-likelihood of these models were *basic compliance*: $-3,097.2$ and *weighted compliance*: $3,832.4$. The results also revealed that the data was severely overdispersed (*basic*: $\chi^2(369, N = 375) = 4,380$ and *weighted*: $\chi^2(369, N = 375) = 2,590$).

Table 4. The coefficient estimates resulting from the negative binomial regression models for two variables of *basic compliance*, and *weighted compliance*.

Groups		Compliance Count				Weighted Compliance			
		β	e^{β}	z	p	β	e^{β}	z	p
Intercept		2.271	9.691	19.054	<.001	1.74	5.67	14.64	<.001
(A) I-PANAS-SF	1. Classic		1.000				1.000		
	2. Chatbot	0.107	1.113	0.658	0.511	0.181	1.198	1.128	0.259
	3. Interactive	0.258	1.295	1.589	0.112	0.303	1.353	1.885	0.059
(B) Affect Grid	1. Classic	0.622	1.862	3.932	<.001	0.695	2.004	4.475	<.001
	2. Chatbot	0.664	1.943	4.178	<.001	0.747	2.110	4.784	<.001
	3. Interactive	0.494	1.639	3.032	0.002	0.636	1.889	3.979	<.001

We then ran separate negative binomial regression models to fit the observed counts of *basic* and *weighted* compliance variables to the regression matrix. To determine the variance in terms of the mean for the negative binomial variance function (α value), we fitted OLS regression models that returned t-scores of the regression coefficient α, showing significant α values for all negative binomial models ($CI = 99\%$, *basic*: $\alpha = .68, t(369) = 12.82$ and *weighted*: $\alpha = .60, t(542) = 12.80$). This indicated that the negative binomial regression model could do a better job of fitting the data than a Poisson regression model. Using the α value, we then used the negative binomial regression models, which showed the goodness of fit with $CI = 99\%$ and *basic*: $\chi^2(369, N = 375) = 436$ and *weighted*: $\chi^2(369, N = 375) = 427$. Table 4 provides the coefficient estimates resulting from these models for each variable. Using the model of *basic* compliance, we observed that compared to users of user group A_1, users of measure group B design alternatives were more likely (B_1: $e^{\beta} = 1.86$, B_2: 1.94 , and B_3: 1.64 times more likely) to answer a sampling event.

Further comparison of user groups (within each measure group A and B) demonstrated that we were unable to find enough evidence to conclude any effect of design alternatives on *basic compliance* within each measure group. However, the expected and observed effect of the design alternatives, although insignificant, was reflected by the coefficient values being slightly in favor of interactive design compared to the classic and chatbot designs in group A, and also for the chatbot designs compared to the classic designs in both groups A and B. However, the interactive design of group B seemed to be less likely to increase compliance compared to either its classic or chatbot design alternatives. The model of *weighted* compliance returned slightly different coefficient values as expected, but the results were generally very similar to those derived from the model of *basic compliance*. Based on this model, compared to users of A_1, users of group B design alternatives were for B_1: $e^{\beta} = 2.004$, B_2 : 2.11 , and B_3: 1.889 times more likely to submit a complete answer.

5.2.2 Completion Rate. To investigate the variable of *completion rate*, we examined the number of completed (versus incomplete) entries, which also carried the total number of sampling events that a user had received ($N_{samplings} = N_{completed} + N_{incomplete}$ and $N_{incomplete} = N_{dismissed} + N_{invalid}$). We considered these numbers as binomial data, i.e. the number of successes per trial. A binomial regression with log-odds link function and the iterative re-weighted least square algorithm was performed ($CI = 99\%$). Table 5 provides a summary of the resulting coefficient estimates, showing significant differences between the user group A_1, with measure group B

Table 5. The coefficient estimates resulting from the binomial regression.

Groups		Completion rate			
		β	e^{β}	z	p
	Intercept	-0.226	0.798	-3.887	<.001
(A) I-PANAS-SF	1. Classic		1.000		
	2. Chatbot	-0.124	0.883	-1.623	0.105
	3. Interactive	0.148	1.159	1.934	0.053
(B) Affect Grid	1. Classic	0.645	1.906	8.752	<.001
	2. Chatbot	0.910	2.458	11.979	<.001
	3. Interactive	0.774	2.168	9.796	<.001

(B_1: $e^{\beta} = 1.91$, B_2: 2.48 , and B_3: 2.17). This means that the chance of having a higher completion rate was better for group B compared to A_1.

We then studied separate parallel models in each measure group to investigate intragroup differences. In measure group A, the likelihood of having a higher completion rate was higher for the interactive design (A_3) compared to the chatbot (A_2), — $A_{3/2}$: $e^{\beta} = 1.31, z = 3.87, p < .001$— however, such a statement was not confirmed for these design alternatives when compared to the classic design (A_1). In group B, chatbot design (B_2) appeared to have a higher completion rate ($B_{2/1}$: $e^{\beta} = 1.30, z = 3.98$) compared to the classic design (B_1), but no other regression coefficients were statistically significant. Comparing the user groups based on their design alternatives (e.g., comparing chatbot designs of A and B), we found that measure group B consistently had a higher completion rate than measure group A for each design alternative ($p < .001$).

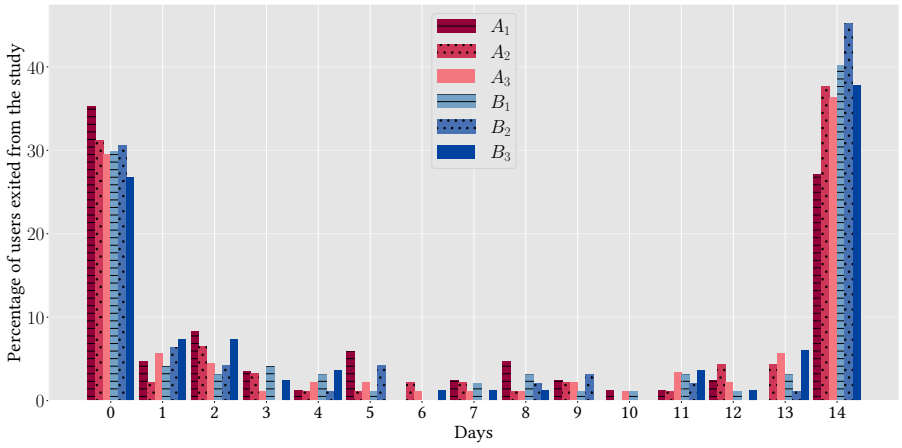


Fig. 4. Percentage of users in each user group who existed the study on each day. The sample is cut-off after the 14th day and includes dropouts within the first day. Group A: the measure of PA-NA dimensions, i.e. I-PANAS-SF and Group B: the measure of pleasantness-energy, i.e. Affect grid; 1: classic, 2: chatbot, 3: interactive.

5.2.3 Dropout. To investigate the variable of *dropouts*, we visualized the exit day from the study in figure 4 considering all participation instances, including those who dropped out within the first day of study (see 5.1.5). The visualization manifests that there is not much of a difference between user groups. Notably, as visualized in figure 4, the majority of dropouts occurred within the first day of the study; and therefore, users' first impression of a measure could be an essential factor in their participation. The percentages of users completing the study

duration of two weeks were as the following: A_1= 41.82%; A_2= 54.69%; A_3= 51.61%; B_1= 57.35%; B_2= 65.12%; B_3= 51.67% when excluding the users dropping out within the first day.

To assess this further, we considered a binary variable as an indicator of those who finished the participation and those who exited the study earlier. The logistic regression model of the variable showed no significant influence of the user groups. Further comparison of the user groups using Kruskal-Wallis H, as well as the post hoc pairwise analysis of all user group pairs using Mann-Whitney U, with days of exit (1, 2, ..., 14) as the variable returned no significant difference between the user groups, either.

5.2.4 Endurance. The *dropout* variable only considered entries and counted days before the 15th days of participation and did not take into consideration that a user may continue using the app. With this regard, we also looked into the *endurance* variable, i.e. the count of days until when a user stopped using the app. The procedure of analysis for this variable was similar to that of *basic* and *weighted* compliance (section 5.2.1). Based on our assessment, the *endurance* variable was over-dispersed, and eventually, we used a negative binomial regression to model it using a log link function ($CI = 99\%$ and $v_5 : \chi^2(369, N = 375) = 388$). Table 6 lists the coefficient estimates resulting from this model. Comparing the resulting coefficients of this model separately with various user groups showed that compared to B_1, A_1 and A_2 user groups had significantly lower chances (almost .5 times) of having a higher endurance.

Table 6. The coefficient estimates resulting from the negative binomial regression model for the variables of *endurance*.

Groups		Endurance			
		β	e^{β}	z	p
	Intercept	4.06	57.985	28.588	<.001
(A) I-PANAS-SF	1. Classic	-0.577	0.562	-2.709	0.007
	2. Chatbot	-0.64	0.527	-3.131	0.002
	3. Interactive	-0.185	0.831	-0.897	0.37
(B) Affect Grid	1. Classic		1		
	2. Chatbot	-0.258	0.773	-1.272	0.204
	3. Interactive	-0.186	0.83	-0.897	0.37

In addition, further analysis showed that users of A_3 were 1.58 times more likely to have higher endurance than users of A_2 ($A_{3/2}$: $e^{\beta} = 1.577, z = 2.177, p = .030$). Finally, users of B_3 had better chances (1.57 times) of having a higher endurance compared to A_2 (B_3/A_2: $e^{\beta} = 1.575, z = 2.151, p = .031$). For other user groups, no more significant effect of user groups was observed.

5.2.5 User Motivation. Variable *motivational* captured user motivation and was calculated by counting the manual entries (i.e. not scheduled). Following a similar procedure to the evaluation of *basic* and *weighted* compliance in section 5.2.1, we ended up with a negative binomial regression ($CI = 99\%$ and $v_6 : \chi^2(369, N = 375) = 394$) to address the slight over-dispersion. Table 7 shows the coefficient estimates of this model. Compared to users of A_1, users of B_1 were $e^{\beta} = 1.78 (z = 2.34, p < .05)$ times more likely to have higher *motivational* value and to submit manual entries.

In addition, users of B_1 also had better chances than users of A_2 ($e^{\beta} = 1.62, z = 2.07, p = .039$) of having higher motivational submissions. However, the obtained p-values here, however, were larger than the adjusted value considered (i.e. .008). Further comparison revealed no additional evidence for other user group pairs.

5.2.6 Time-Spent. Measure groups A and B had different lengths and, accordingly, response durations. A_1, A_2, and A_3 measurement tools were longer and had ten questions, whereas B_1, B_2, and B_3 had only one or two questions depending on the design alternative. Users, therefore, spent more time completing questions of measure

Table 7. The results of the negative binomial regression model for the variables of *motivational* responses.

Groups		Motivational			
		β	e^{β}	z	p
	Intercept	0.656	1.927	3.488	<.001
(A) I-PANAS-SF	1. Classic		1.000		
	2. Chatbot	0.090	1.094	0.354	0.723
	3. Interactive	0.370	1.448	1.461	0.144
(B) Affect Grid	1. Classic	0.575	1.778	2.336	0.019
	2. Chatbot	0.223	1.25	0.887	0.375
	3. Interactive	0.481	1.617	1.892	0.058

group A compared to measure group B (note the median values of the response duration in table 3). One may argue that in addition to the longitudinal commitment of the users, the number of questions they answer could be an indicator of their compliance. To count for such an argument, we explored the total time spent with each tracking tool in hours by each user. As the summary of variable *total time* in table 3 displays, B_2 followed by B_3 had the highest average time-spent per user group. Surprisingly, shorter measures, such as those of measure group B, had a higher *total time-spent* value compared to measure group A. This means that not only would the users continue using a shorter measure, but they would also spend more time overall with those measures as a result.

5.2.7 Investigating the Interaction Effects. As we saw throughout this section, the majority of the variables we introduced for capturing user compliance in this paper were count variables. Accordingly, it was not possible to investigate these variables using analysis of variables right away. Although we investigated all variables thoroughly with proper models, to consider the possible interaction effect further, we additionally ran a two-way multivariate analysis of variance after transforming the count data with a log transform function. We also excluded the *basic compliance* variable due to its strong significant ($p < .001$) correlation with the *weighted compliance* variable ($r = .981$). We found a statistically significant interaction effect between the type of measure and design alternatives on the combined dependent variables ($F(12, 508) = 2.635, p = .002, \text{Wilks}'\Lambda = .886, \eta^2 = .059$). Follow-up univariate two-way analysis of variance showed a statistically significant interaction effect between the type of measure and design alternatives for variables of *completion rate*, *motivational responses*, and *total time-spent* with the tool, but not for *weighted compliance*, *dropout rate*, or *endurance*. Table 8 displays the obtained values from the analysis.

Table 8. Two-way analysis of variance showing the interaction effect between the type of measure and design alternatives.

	Variable	F	p	η^2
measure * design	Weighted compliance	2.268	0.106	0.017
	Completion rate	3.82	**0.023**	0.029
	Dropouts	0.962	0.383	0.007
	Endurance	0.942	0.391	0.007
	Motivational	6.602	**0.002**	0.049
	Total time-spent	3.447	**0.033**	0.026

As such, a simple main effects analysis was conducted for *completion rate*, *motivational responses*, and *total time-spent*. Table 9 lists the results of this statistical evaluation, highlighting the statistically significant difference between measure A and B using classic and chatbot designs on completion rate, but not for the interactive design. There was also a statistically significant difference between measure A and B using classic designs on

Table 9. Main effects analysis of design for variables of completion rate, motivational responses, and total time spent.

Variable	Design	df	F	p	η^2
Completion rate	1. Classic	259	6.729	**0.010**	0.025
	2. Chatbot		27.451	**<.001**	0.096
	3. Interactive		1.311	0.253	0.005
Motivational responses	1. Classic		18.697	**<.001**	0.067
	2. Chatbot		0.005	0.946	<.001
	3. Interactive		0.033	0.855	<.001
Total time-spent	1. Classic		0.032	0.859	<.001
	2. Chatbot		13.111	**<.001**	0.048
	3. Interactive		1.342	0.248	0.005

motivational responses but not for the chatbot or interactive design. In the total time spent with the app, there was a statistically significant difference between measure A and B using chatbot designs, but not for the classic or interactive design.

Table 10 lists the results of assessing the main effect of the measure. There was a statistically significant difference between design alternatives for measure A for completion rate but not for measure B. Also, for motivational responses, there was a statistically significant difference between design alternatives for measure A but not for measure B. The difference between design alternatives in total time spent for both measure A and B was non-significant. Additional pairwise comparisons revealed no additional information than those obtained earlier from the regression models.

Table 10. Main effect analysis of measure for variables of completion rate, motivational responses, and total time spent.

Dependent Variable	Measure	df	F	p	η^2
Completion rate	(A) I-PANAS-SF	2	4.785	**0.009**	0.036
	(B) Affect Grid		1.898	0.152	0.014
Motivational responses	(A) I-PANAS-SF		4.914	**0.008**	0.037
	(B) Affect Grid		2.174	0.116	0.017
Total time-spent	(A) I-PANAS-SF		2.247	0.108	0.017
	(B) Affect Grid		1.243	0.29	0.01

5.2.8 User Self-reflection. Users of measure group B exhibited higher compliance as a whole over time, which may at first appear to be related to the response duration of the measure. To evaluate the shortness of the measures and their design alternatives further in this experiment, we used an enhancement of the Keystroke-Level Model [12] for mobile devices [22, 49]. Using Touch-Level Model (TLM) [49], we then modeled each measure and design alternative (or tracking tool) and calculated an approximate time for completing a sampling with each one. To have a comparison between the tools, we uniformly used the following operators [22, 49]: Mental Act (1.35 s), Tap (.08 s), and Swipe (.07 s), and considered a mental operator for each question item or label. The resulting values in second were: $TLM(A_1) = 18.73$; $TLM(A_2) = 18.43$; $TLM(A_3) = 15.63$; $TLM(B_1) = 12.31$; $TLM(B_2) = 5.64$; $TLM(B_3) = 4.27$. These results are mainly in accordance with the response duration values in table 3 if we consider the time that a user spent on self-reflection for answering the questions as the difference between median-response duration and TLM (i.e. $A_1 : 34.27; A_2 : 21.57; A_3 : 16.37; B_1 : 7.69; B_2 : 13.36; B_3 : 9.73$ seconds). There was also a strong and significant correlation between the calculated TLM scores and the median response duration for the tools ($r = .895, p < .016$)— additionally supporting the strength of TLM in determining the

response duration in mobile tracking. Users seem to spend less time for self-reflection in B_1 compared to other user groups. This observation is reasonable considering that while users of measure A have to think about at least ten different affective states and determine their extent according to how they have been feeling for the past few hours, users of B have to just specify the pleasantness and energy level of their feelings providing either one (B_1) or two (B_2, B_3) responses.

5.3 Discussion: RQ2-user Compliance

5.3.1 Weighted and Basic Compliance. We investigated seven variables of compliance and compared six user groups, i.e. two measures and three designs each, in this part. Assessing the *basic* and *weighted* compliance variables showed that the measure, but not its design, had a significant impact on user compliance. It also points out that a shorter measure in length, i.e. fewer questions, may not necessarily lead to higher compliance. For example, none of the measures of group B with only two questions statistically outperformed the interactive version of measure A with ten questions. This observation also suggests the possible positive effect of design in engaging users in the studies. Interestingly, *weighted compliance* strengthened the coefficient values of the shorter measures in spite of more taxation of the compliance values in measure group B compared to A; e.g., the likelihood of a higher compliance value of $B_{2.chatbot}$ compared to $A_{1.\ classic}$ increased from 1.94 times (in *basic compliance*) to 2.11 times (in *weighted compliance*).

Nevertheless, *weighted compliance* values still resulted in the same outcome as the *basic compliance* values, and no other impact, other than that of *basic compliance*, was observed using *weighted compliance*. Considering the strong correlation between the basic and weighted variables, we can also conclude that taxing responses by response delay or duration has no major impact on the outcomes of user compliance in our study as a whole. This could be related to the design of our ESM, which had a dismiss button for a sampling event, enabling users to dismiss the sampling prompt immediately. It is, therefore, likely that a user with lower motivation dismisses the event instead of responding too quickly, resulting in close values of basic and weighted compliance variables in our evaluation. We also want to point out that as we saw earlier in table 3, both mean and median values for weighted compliance were much closer to the basic compliance for measure group A, compared to B. Therefore, the weighted compliance value still could be a good solution, especially for cases where there is no option for dismissal of the sampling event or when using shorter measures, and therefore, it should be further explored.

5.3.2 Completion Rate. Analyzing the *completion variable* introduced partly contrary results compared to the *basic* and *weighted* compliance variables. Similar to the *basic* and *weighted* compliance variables, the measure type had a significant influence on the *completion* variable. Nevertheless, the *completion* variable also showed the influence of design alternatives for a measure group. There was also a significant interaction effect between the measure and design alternative. Depending on the measure, at least one of the design alternatives of chatbot or interactive could potentially improve the completion rate of a measure compared to its classic design. In addition, the coefficients of B_1 for the *basic* and *weighted* compliance variables were higher than those of B_3, whereas, for the *completion* variable, it was the other way around. Since this variable gives a more realistic perspective of user compliance, we encourage the scientific community to carefully consider the presented sampling events in calculating compliance and count for events that participants may not have received.

5.3.3 Other Variables. Other variables of this experiment revealed additional information about the user groups. We found that the user group had no significant impact on the *dropout*. Nevertheless, the percentages of users who completed the study duration were, as a whole higher for measure group B; yet, except for the chatbot design, these percentages were not consistently higher or lower for a specific design alternative. Although only slightly, the chatbot design seemed to have the highest percentage of users completing the study duration in each measure group (had the lowest dropouts). Users of B_1 had higher endurance, and they were more likely to

keep using the app for a longer time than users of A_1 and A_2. This was the case for users of A_3 compared to A_2, as well. Users of B_1 were also more likely to submit manual entries than users of A_1 and A_2.

To sum up, whether an interactive, chatbot, or classic design is better for user compliance, on the whole, depends on the measure. For measure group A, while the chatbot and classic designs were very similar in compliance behavior and motivation (the *basic* and *weighted* compliance and *motivational response* variables) of a user, the chatbot had more dismissed entries and less response duration on average. The interactive design, however, clearly outperformed the chatbot (but not the classic) design of this group in *completion rate* and *endurance*. In contrast, for measure group B, it was the chatbot that had a significantly higher chance of a better *completion rate* than its classic counterpart. For other variables, although the mean or median values of a design alternative were lower or higher than another design alternative within measure group A or B (table 3), their differences were non-significant. Altogether, in group A, the interactive design, and in group B, the classic design seems to be slightly better than other user groups over time when we take all variables into consideration. It is also important to note that higher user compliance does not necessarily mean better user engagement, as it may be accompanied by clumsy or meaningless user inputs instead of real answers in the long run.

5.3.4 User Self-reflection. Considering the lengths, response duration, and TLM of the measures and their design alternatives, one may, at first glance, argue that compliance is only a factor of the measure's length; however, that is *not always* the case. For instance, in several variables (e.g., basic and weighted compliance, endurance, and motivational responses), B_1 appeared to outperform other tools, but this was not consistently the case for its design alternatives, such as B_2 or B_3 that were all very short measures. In fact, there was no evidence of the interactive design of group B (B_3) significantly outperforming A_3 user group in the basic compliance, weighted compliance, dropout, endurance, and motivational responses variables or any other user groups in the total time spent variables. Therefore, the length of a measure per se cannot predict user compliance in all variables, and other factors, such as user experience, may also be involved. Its impact could, however, appear in at least one of the compliance variables.

On average, users of B_1 and B_2 had the most number of completed events (scheduled and manual) per user, compared to other user groups, but they potentially could have submitted more meaningless responses, as well. Dividing the total number of completed events by the number of users of each group in table 3 yields $A_1 : 24.65; A_2 : 29.04; A_3 : 30.32; B_1 : 34.07; B_2 : 31.63; B_3 : 29.15$, which shows the average number of completed events per user. Accordingly, users of B_1 have completed more sampling events on average. Besides, the user effort of submitting a sampling event in measure group B was similar to dismissing it since users record their mood state with only 2–3 tap or swipe operators (according to TLM). However, completing a sampling event with design alternatives of this measure group also had several Mental Act operators. If users of measure group A wanted to submit meaningless responses, they would face at least 11–15 Tap or Swipe operators, compared to two tap operators, to cancel (or dismiss) the sampling event. Consequently, it is more plausible for a user of measure group B to just submit a meaningless response instead of either thinking and self-reflecting or dismissing the sampling event. This argument is well-supported by users of B_1 spending less time on average on self-reflection, according to TLM and response duration values. It also is supported by the higher decrease in the median and mean values of weighted compliance — which taxes responses that were submitted too quickly — compared to the basic compliance for group A compared to B according to table 3. Both the low response duration and low TLM values in measure group B, together with the design of this measure and the total number of submissions, imply that very short measures can potentially also attract more meaningless responses. In other words, a user may just randomly select a value instead of dismissing the sampling event to get rid of it quickly. To complete the comparison between the measures and their design alternatives, we next look into the user experience of each user group.

6 PART 3: RQ3-USER EXPERIENCE

The aim of the third part was to compare the user experience of our six mood tracking tools (i.e. 2 mood measures×3 design alternatives). We hypothesized that the design alternatives of a measure would have better usability and user experience scores than their classic counterparts. This part is the result of the post-study survey.

6.1 Method: RQ3-user Experience

After two weeks of ESM-based study described in RQ2 (section 5.2), the app automatically promoted a link to an online post-study survey using the Sosci platform [37], where users answered the user experience questions. Since the longitudinal study had a between-subjects design, correspondingly, so did the post-study survey following up on the user experience of the mood trackers. In total, 156 participants completed the survey following the previous part (section 5). To ensure that the participants had used the app before the evaluation, we filtered them out based on their participation in the previous part in which they used the app everyday (described in section 5.2, and specifically, subsection 5.1.5). We also removed self-claimed meaningless entries. In the end, 122 participants (female:83, male: 37, other options: 2) remained with an average age of $M = 33.11, SD = 12.53$ years old. The number of users in each user group was as follows: $A_1 = 13, A_2 = 18, A_3 = 19, B_1 = 24, B_2 = 30, B_3 = 18$.

6.1.1 Post-study Survey Structure. The survey had the following elements: (i) Open-ended questions on what users liked and disliked using the app; (ii) User Experience Questionnaire (UEQ) [36] (internal reliability Cronbach's $\alpha = .95$); (iii) System Usability Scale (SUS) [10, 11] (internal reliability Cronbach's $\alpha = .96$); (iv) Modified questions[7], reluctance and meaningless response questions, briefing, and debriefing.

While most questions of this post-study survey had a Likert-type scale response format, element (i) asked open-ended questions. In this element, users described what they liked and disliked about the app in two separate questions. There were no restrictions to force responses of a specific length; therefore, users were free to respond to these questions in writing. We conducted a content analysis on these written answers and identified the major viewpoints. The quotes of each participant are marked with the participant's number (e.g., P_{13}).

6.2 Results: RQ3-user Experience

6.2.1 Users' Opinions of What They Liked. The open-ended questions of the survey collected the user statements. Each of these statements reflects the user's experience and perception of the app and, correspondingly, the respective assigned measure and design alternative to the user, which has been used at least twice daily. Some of the identified qualities were mainly related to the app design, e.g., features of the app mentioned by users, and, therefore, are not exclusive to a specific mood tracker tool and are shared between all user groups. For example, several users liked the regular pop-up function of the app that reminded them to record their mood ($n = 12$, e.g., $P_{4,9,84,102,etc.}$). Namely, P_{27} reported: "*I liked how it popped right up on my phone, so I could not forget about it…*". Users also generally appreciated the feeling of self-awareness and engaging in self-reflection ($n = 15$, e.g., $P_{26,35,100,122,etc.}$). For example, P_{120} stated: "*I liked the possibility of self-reflection…*". P_{35} commented: "*I liked that you see which days you are happier or not and then you think about what the reason could be*". P_{83} wrote: "*I liked that you think about how you actually feel and what influences that*".

Other qualities that users wrote about were more specific to their assigned user groups and revealed what users noticed and liked about their measurement tool the most. Users of A_1 thought that the questions could be answered easily ($P_{2,4,6}$) and quickly (P_4). Users of A_2 liked the chat appearance and the self-reflection ($P_{24,26}$).

[7] In addition to SUS [10, 11] (which contains the following items: interest in frequent use, complexity, ease of use, need for support to use, functions being well-integrated, inconsistency, learnability, cumbersomeness to use, confidence in use, need to learn before use) and UEQ [36], we included a set of modified questions specifically designed to capture burden aspects at the end of the survey. However, since the analysis of these questions did not reveal any additional meaningful difference between the tools, we excluded them from this investigation for brevity.

P_{26} commented: "*... you are stimulated to think about your own mood ...*". Overall, users of A_3 provided more description about what they liked compared to users of A_1 and A_2. They considered the A_3 questions to be thorough and clear to answer ($P_{34,43}$). They also appreciated the feedback functionality of the app and recalled their measurement tool as simple (P_{41}), easy to use ($P_{39,47}$), quick ($P_{49,50}$), and well-designed (P_{40}), e.g., P_{39} reported: "*I enjoyed how simple it is to use ...*". P_{48} pointed out: "*... it helped with my self-esteem*".

Users of B_1 praised the idea of the app (P_{51}) and the interface ($P_{58,62}$). They also saw this measure as simple ($P_{64,66,67}$) and quick ($P_{65,70}$), e.g., P_{64} wrote: "*... Probably among the simplest interfaces I've used ...*". Users of B_2 also considered their measure easy to use (P_{78}), simple ($P_{81,82}$), quick ($P_{82,102}$) and enjoyed the chatbot-form interaction (P_{85}) and its visual design (P_{80}). B_3 was quick ($P_{106,110,115}$), easy to use ($P_{116,119}$), uncomplicated to operate (P_{116}), and simple ($P_{108,110}$) in its users' opinion.

6.2.2 Users' Opinions of What They Disliked. Users of measure A were mostly dissatisfied with items or affective states with which they had to specify their mood. What users did not like about A_1 was mainly related to the layout (P_8) as well as the meaning and coverage of its item ($P_{3,4,5}$), e.g., P_{11} stated: "*For me, having hostile ... was not accurate or helpful*". The same goes for users of A_2 ($P_{20,26}$) and A_3 ($P_{35,43,48}$). P_{20} reported: "*The questions that were asked were limiting ...*". P_{35}, a user of A_3 commented: "*For me, the mood items are not enough or should have been different ... I had difficulty deciding ...*". From such statements, it seems that some users started thinking about their feelings with the app and then looked for more specific mood states to express their feelings accurately.

Users of B_1 thought that the app had too many graphics (P_{52}), was too brief ($P_{52,66}$), and did not ask enough or proper questions ($P_{52,55,58,61,65,74}$). For example, P_{71} reported: "*The moods and causes were not scrutinized enough*". A few users also considered this tool difficult ($P_{54,73}$); P_{73} wrote: "*... it was hard to determine ... and indicate my mood on the coordinate system ...*". Users of B_2 found it to be too short (P), limiting (P_{100}), difficult ($P_{83,87,97}$), and to have too few options ($P_{77,83,102}$). P_{79} found the interface boring. Users of B_3 believed that they could not reflect their mood using manikins intuitively and with enough details ($P_{109,116,119,120,121}$). P_{105} disliked that the same questions were shown over and over.

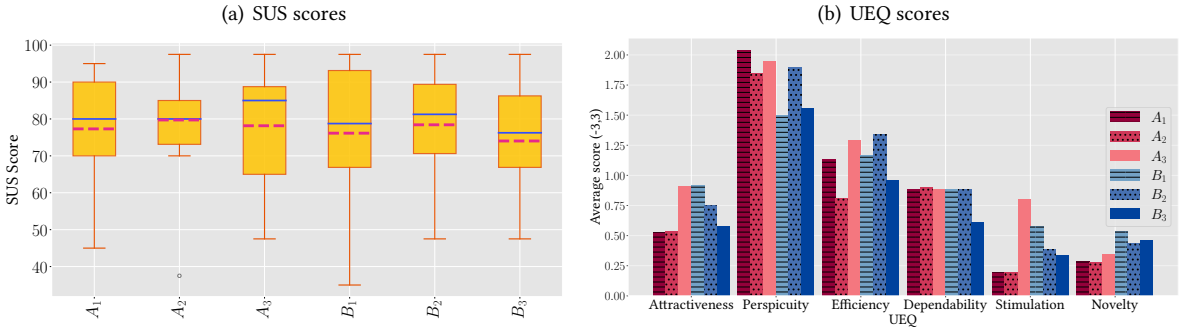


Fig. 5. (a). Box plot of SUS scores for each user group. The straight blue line is the median, and the dashed pink line shows the mean values. (b). Bar plot of the UEQ scores for each user group. UEQ scores are between [-3, 3]. A: I-PANAS-SF capturing PA-NA dimensions; B: Affect Grid representing pleasantness-energy dimensions; 1: classic; 2: chatbot; 3: interactive.

6.2.3 Quantitative Evaluation. Two-way multivariate analyses of variance on obtained SUS and UEQ responses with two independent variables of measure and design were run. The interaction effect between measure and design on the dependent variables was non-significant $F(14, 220) = 0.922, p = .536$, Wilks' $\Lambda = .892, \eta^2 = .055$. The main effects for both measures and designs were also non-significant. Accordingly, no evidence of significantly

better scores — that were given by users — for any of the tested variables of UEQ or SUS was found. Figure 4(a) visualizes the summary of SUS scores for each user group. The SUS scores for all user groups were pretty similar, good, and close to excellent, according to the Bangor et al. [4]'s definition. Specifically, scores of A_1 — i.e. classic I-PANAS-SF — ($M = 77.308, SD = 14.558, Mdn = 90$) and B_1 — i.e. classic Affect Grid— ($M = 76.146, SD = 19.125, Mdn = 78.750$) can serve as a baseline for future improvements. Figure 4(b) depicts the average UEQ scores for each user group. Similar to the SUS scores, no values of UEQ were significantly different between user groups.

6.2.4 User Experience and Compliance Variables. In section 5, we thoroughly investigated user compliance with several variables and found interaction effects between the type of measure and design alternatives for the completion rate variable. To explore the user experience of a mood tracking tool concerning its completion rate, we built a Generalized Linear Model using Binomial family and logit link function. As mentioned before, the completion variable counts the number of completed sampling events from the presented ones to the users, assuring that the event actually has been received by the user based on the log data. Therefore, all failed sampling events were excluded, resulting in a varying number of presented events per user and making the data as success per trial type. For this model, we excluded the attractiveness score of UEQ due to its high correlation with other factors (a correlation of $r = .749$ with efficiency and $r = .86$ with stimulation). Dependability was also excluded since it was almost equal for all tools. In this model, we used TLM (as discussed in section 5.2.8) instead of measure or design alternatives since it represents the differences between the tools, and it is easier to comprehend in the model.

The resulting model ($\chi^2(6, N = 118) = 580, p < .001$) for predicting the probability of an event completion was as follows: For a given record, the probability of sampling event completion is $-.380 + \text{TLM} \times -.067 + \text{SUS} \times .026 + \text{UEQ.Efficiency} \times .224 + \text{UEQ.Stimulation} \times -.163 + \text{UEQ.Novelty} \times -.098 + \text{UEQ.Perspicuity} \times -.092$. This means that based on this model, a successful outcome, i.e. completion of a sampling event, has an increased chance (1.251 or 25.1% if other predictor variables are kept at a fixed value) for every unit of increase in the UEQ-Efficiency score. The odds of completing an event slightly increase ($e^{.026} = 1.027, p < .001$) for every unit increase of the SUS score, but it decreases ($e^{-.067} = .935, p < .001$) for every unit increase in TLM. The probability of having a sampling event completed also decreases with a unit increase in UEQ-Stimulation ($e^{-.163} = .850, p < .001$), UEQ-Novelty ($e^{-.098} = .907, p = .033$), and UEQ-Perspicuity ($e^{-.092} = .912, p = .045$).

We also modeled other compliance variables discussed in the previous section (section 5.2). In particular, a negative binomial regression model with the endurance variable ($4.113 + \text{UEQ.Stimulation} \times .182$) suggested that for every unit of increase in UEQ-Stimulation score, the chances of having a higher endurance increases by about 19.9% ($e^{.182} = 1.199, p = .015$). A Poisson model for motivational responses was used, and accordingly ($.851 + \text{TLM} \times -.055 + \text{SUS} \times .016 + \text{UEQ.Efficiency} \times -.322 + \text{UEQ.Stimulation} \times .186 + \text{UEQ.Novelty} \times -.158$), for every unit of improvement in UEQ-Efficiency score, the chances of more motivational responses decrease $e^{-.322} = .725$ ($p < .001$) times while keeping TLM, SUS, UEQ-Stimulation, and UEQ-Novelty at a fixed value, and similarly for TLM ($e^{-.055} = .946, p < .001$), and UEQ-Novelty ($e^{-.158} = .853, p = .025$). For every unit of improvement in SUS, the chances of having motivational responses remain almost the same with only 1.62% improvement ($e^{.016} = 1.016, p = .001$). This improvement is about 20% when increasing UEQ-Stimulation by one unit ($e^{.186} = 1.205, p = .006$). No other additional insights were obtained from the models.

6.3 Discussion: RQ3-user Experience

Comparing user groups with various qualitative and quantitative user experience assessments revealed the strengths and drawbacks of each measure and its design alternatives. As a whole, users of measure group A (I-PANAS-SF) considered their measure to be easy, thorough, and clear to answer but were critical of having specific mood states. In contrast, users of measure group B (Affect Grid) regarded their measure as quick and were fascinated by the visual elements of its design alternatives overall. The results of the quantitative assessment

altogether support these qualitative outcomes. For example, in the qualitative assessments, users of measure group A did not discuss the visual elements as much as users of measure group B, and correspondingly, the attractiveness and stimulation scores of group A, generally and on average, were lower than group B — except for the interactive design of group A.

Taking all scores of SUS, UEQ, and modified questions into account, we found that with our app design, there was no significant difference between user groups; and accordingly, our initial hypothesis did not hold due to several possible reasons as follows. The user groups' population was not equally distributed, and fewer users were involved in some user groups than others, possibly because of users' fatigue over time. In addition, there was no within-subjects baseline to compare an individual's evaluation of the app's user experience or usability with such a baseline. This mainly occurred since our study design required users to interact with a tool for a more extended period of time before participating in the post-study survey (i.e. the third part). We chose to do that in order to capture the user experience of the mood measurement tools for experience sampling in a longitudinal study (i.e. mood tracking). It is plausible that the users' first impressions of a measurement tool differ from their impressions after repeated use. For example, we observed in the first part (section 4.2.6) that overall, users disliked the classic version of Affect Grid (B_1) the most. However, this tool seems to be comparable with other tools in user experience evaluation after two weeks of use, as well as in user compliance. Including any arbitrary baseline designs in the post-study survey without repeated use, therefore, could have had an influence on the evaluation of the used measure. On the whole, it was not conveniently possible to include another measure as a baseline and compare it with users' assigned mood tracking tool, within-subjects and without any bias.

Another reason could be related to the design of the app. The results were generally consistent with the possibility that the user experience of a measure depends more on the overall design of the app rather than the design of the measure. In the qualitative assessment, several users of various user groups similarly mentioned the app's impact on self-awareness and self-reflection, and they discussed the timing of the notification, feedback functions, etc. The design and functionalities of the app, which were kept the same for all user groups, could therefore have had a more substantial effect on users' impression of the app and, accordingly, caused the small observed differences in user experience among user groups. This possibility highlights the importance of using a properly designed measurement tool with smartphones and its impacts on attracting users and achieving a high level of user experience more than ever. The design alternatives of each measure group were not consistently or significantly better or worse than other design alternatives. Overall, a well-designed app would have good and close to excellent usability regardless of the measurement tool used.

The general presumption and attitude toward mood tracking in particular, and ESM in general, is that the shorter the measures, the better it is for repeated assessment. However, our comparative study demonstrates that other critical criteria, such as the measure itself and its user experience, are more influential determinants of user commitment in tested measures of this study. Modeling various compliance variables based on the obtained TLM and user experience questionnaires using generalized linear models provided additional insights into the characteristics that are important in increasing the likelihood of more user compliance. Namely, improving the efficiency as captured by UEQ of a design alternative for a measure, which is about enabling users to accomplish the task without unnecessary effort, can improve the completion rate overall up to 25%. While efficiency seems to be a relatively effective and important criterion, TLM, or the measure's length, can improve compliance by only about 6.5%.

We also found that various aspects of user experience have possibly varying impacts. SUS, based on our models, could, at best, improve the compliance variables of completion rate and motivational responses only about 2%. However, this improvement was different depending on the factor of UEQ. UEQ-Dependability was almost the same for all tested tools. Attractiveness was also excluded due to its strong correlations with efficiency and stimulation. This also suggests that for a mood tracking tool, its efficiency and perspicuity (i.e. easy to get familiar with) are significantly associated with the overall user impression of the app or its attractiveness.

Efficacy, as we saw, was the strongest variable in our models for increasing the likelihood of a better completion rate. Nevertheless, when considering motivational responses, this factor may negatively impact the chances. In particular, the odds of motivational responses may decrease about 27.6% with every unit of increase in this variable. We observed a similar but reverse pattern for stimulation. While the more stimulation a tool has, the chances of submitting self-motivated responses increase about 20% for every unit improvement in stimulation; the odds of completing a sampling event reduce by about 15%. In short, an increase in efficiency or decrease in stimulation based on our models may increase system prompt sampling completion and reduce self-motivated responses. These quantitative outcomes provide a clear path on the decision while choosing a design or measure. When the study's intention is oriented toward system prompt samplings, efficiency should be the target for improvement. However, the higher the stimulation and more exciting the tool is, it would attract more self-motivated responses and increases the likelihood of more user endurance.

7 OVERALL DISCUSSION: IMPLICATIONS

Taking together the results of the study in three parts (RQ1, RQ2, and RQ3), we found that the type of mood measure and its interaction design impact the validity and quality of the self-reported mood measure. Various measures have different resilience toward design changes, and while one can be very resilient toward changes, another would not be valid anymore given even the slightest modifications. We also saw that different measures and design alternatives have varying user compliance over time. However, the measure or its design does not significantly influence the perception of the app's user experience. These observations have several implications for conducting research related to mood, health, and behavior tracking, as well as assessing user experience and compliance in ESM studies.

7.1 A Resilient Measure

When a mood measure is resilient toward changes, it seems that its interactive design could be more attractive in one-time assessment and user compliance, as well as in user experience assessments. This was the case for I-PANAS-SF. This measure generally is very resilient toward chances. The concurrent assessment of it as measure group A showed that the chatbot, classic, or interactive design alternatives could be used safely for capturing the mood variables of PA and NA.

Users liked the interactive design better in their first impressions, and after two weeks of use, they also scored its usability on average slightly better compared to other design alternatives scored by their users (though the differences were non-significant). This design had the lowest TLM value and median response duration in group A, as well. However, users also disliked this design more than its design alternatives in their first impressions (section 4.2.6). In other words, this design was more controversial since, as a whole, users both liked and disliked it the most compared to its design alternatives. This design, in the end, had a significantly better user completion rate and endurance than its chatbot counterpart.

The chatbot design of measure group A was both liked and disliked the least in the first part (first impressions of users in section 4.2.6) and scored the least in usability evaluation by its users on average, as well. This design also had the least average user-reported efficiency between all user groups and perspicuity in measure group A. Its TLM and median response duration were slightly lower than the classic design. In other words, this design was less controversial as a whole and had less usability, particularly efficiency. In the end, although its completion rate and endurance were significantly lower than the interactive design, it was very similar to the classic design for any other compliance variables in general.

Taken together, the completion rate and endurance in this measure group could depend generally on the likeability and user experience, and particularly efficiency, stimulation, and perspicuity, of the design. This

could explain the difference between the chatbot and interactive design of this measure group and the higher compliance (completion and endurance) of the interactive design.

7.2 A Brief but Non-resilient Measure

The classic Affect Grid is generally susceptible to modifications. We saw that it had a low agreement with its design alternatives. Users seemed to dislike this measure the most at the beginning. In the end, although many users struggled with recording their mood with this measure, even after two weeks found it difficult, or thought that they could not reflect their mood states with it, they gave it good overall usability scores after repeated use. This measure also appeared to function better in user compliance compared to the I-PANAS-SF. However, a drawback of Affect Grid is the similar cost of dismissing a sampling event against submitting a meaningless response for a user, which could potentially lead to a good score of compliance but inaccurate measurements. This measure could still be a good measure of *emotions* if users are instructed and trained carefully according to the original instructions in [55]. Nevertheless, the outcomes of this study as a whole raise caution regarding the validity of this measure for repeated assessment of mood for a remote and longitudinal smartphone-based study.

In contrast to the classic Affect Grid, its interactive design, for pleasantness dimensions, appears to be more compliant to the theories since it showed almost a perfect relationship with PA-NA dimensions as one expects (discussed in section 4.2.5). This design was *liked* slightly more in users' first impressions and *disliked* a lot less than the classic Affect Grid (discussed in section 4.2.6). However, it did not significantly perform better than the classic version in user motivation, endurance, or compliance, or completion.

After two weeks of usage, the chatbot version of group B had the best average user-perceived usability among its design alternatives, while the interactive version had the worst (note that the differences between user groups for user-perceived usability were non-significant). The chatbot design was both liked and disliked the least in group B by the users in the first part (first impression in section 4.2.6), yet it outperformed its classic counterpart in completion rate. Therefore, unlike in measure group A, in this measure group, we cannot reach a unanimous conclusion about the superiority of one design over another. In other words, depending on the variable of interest, one design is preferable over another. It is noteworthy that the perspicuity and efficiency scores of the chatbot design were slightly higher than its classic design on average. The observation, that in both groups A and B, the design alternatives with a significantly higher completion rate than their design alternative counterparts had slightly higher perspicuity and efficiency scores, suggests that the association of these variables should be further investigated. Measure group B overall functions well when user compliance is a more important target than the accuracy or quality of the mood measurements.

7.3 A Shorter Measurement Tool Does Not Always Bring Higher User Compliance

We also observed from the qualitative results that sometimes brief measures can add more complexity to the task and leave users unsatisfied with limited options to express themselves, such as in the interactive design of Affect Grid (B_3). In addition, we observed earlier that the interactive design of measure group B (Affect Grid), despite its minimum TLM value between all mood tracking tools, did not outperform the interactive design of group A (I-PANAS-SF) — which had more than three times more TLM value — in user compliance. In other words, a measure with only two questions had one-fourth TLM and half response duration median value compared to another measure with ten questions, could not outperform the longer measure, which scores better in usability evaluation after repeated use, in basic or weighted user compliance variables. The same applies to the chatbot or interactive versions of measure B compared to its classic version. These findings suggest that user compliance is not just about response duration, TLM value, or shortness of a measure. Altogether, while selecting a design or measure, user experience-based metrics should have a higher priority than the measure's length in decision-making.

7.4 Which Measure? Key Recommendations and Takeaways

Ultimately, choosing a measure or modifying it would be a three-way trade-off between user compliance, accuracy, and user experience. In essence, a mood measure should first be chosen based on the requirement of the study and the underlying affect theory to which the factors under study belong. In addition, often, the population under study is an important factor in choosing the proper measure. For example, there are specific measurement tools that are deemed to be more suitable for specific user groups, e.g., with specific clinical conditions (see [5, 52, 56, 70]). The measurement tools should be selected and used based on their relevance and compatibility with the study, considering the conditions for which they have been tested and deemed reliable, such as both the population description and measurement concept or theories, as well as the tool's psychometric properties. Accordingly, aspects such as the scope, mood versus emotion, affective states, stimuli, population characteristics, and the context of use should be clarified from the beginning.

Given the interdisciplinary character of affect research, researchers with diverse backgrounds tend to solely emphasize the aspects that are more valuable to them, which might be a source of confusion in interdisciplinary research. In the example of a mood tracking app using smartphones, it is easy to find solutions from the psychology or technical community with a poor interaction design (poor app quality). They may, therefore, have unpredicted, hidden, or even invalid results from the users as a possible outcome of the poor interaction design. The importance of app quality in study design becomes more apparent after the study described in this paper. Likewise, it is simple to find smartphone apps with both convenient interaction design and good technical foundation but with a considerably inferior basis in affect science (poor assessment quality). These applications can, therefore, be unreliable and invalid in capturing mood values from a psychological perspective. Typical examples of such apps are those studies that use arbitrary questions and response formats or modify a non-resilient measure without any validation.

A successful mood tracking app should consider at least two main facets: (1) The assessment quality (including using the proper measure and the validity, reliability, and comparability of the used measure with previous studies) and (2) the app quality (including user experience of a measure, as well as its functional quality, which altogether can also influence the quality of the measurements and user compliance over time). This advice encourages the scientific community to close the gap between fields in an interdisciplinary research area, such as affect research.

When choosing a general-purpose mood measurement tool for non-clinical populations, and when the tracking of user mood or its fluctuation is of concern, classic design for I-PANAS-SF can achieve a high level of accuracy and capture valid responses. Although the general assumption is that such a measure may be boring or disengaging for users, our study suggests that it is not always the case, and many users still prefer the classic design when the app is well-designed (e.g., excellent usability, see section 6.2.3 and [4]). A well-designed app with an interactive design can engage users even more.

A critical finding of this study is that converting scores of one measure to another, particularly in measures with low resilience toward changes as well as in measures of different theoretical backgrounds, such as PA-NA and pleasantness-energy dimensions, should be avoided. A long-term trend or model that has been achieved with lower accuracy of measurement cannot be trusted in critical domains, such as health, where accurate measurement is of concern. What one chooses for one study may not apply to another, and this draws special attention to the importance of a proper measurement tool selection and mood assessment.

Taken together, it is essential for every study using mood measurements, and particularly mood tracking, to determine and discuss the resilience of the applied measure. In case the selected measure was not resilient enough, one should either avoid modifying it or validate the modified tool. Otherwise, it would result in poor assessment quality. In addition, for any newly developed or modified measurement tool, as well as any claims of improvements, user compliance and usability of the presented tool should be clearly discussed and compared against the existing measurement tools, such as those presented in this paper. When user compliance is of

concern, more attention should be paid to the app design and usability, while preserving the assessment quality. In particular, stimulation, efficacy, novelty, and perspicuity of the designed tool are factors to monitor closely in design of the measurement tools. As this paper demonstrated, for the case of the two tested measures, user experience of the tools seems to be more influential than the measures' length — of two versus ten questions — in determining overall user compliance.

7.5 Limitations

In this paper, we compared two mood measures with three design alternatives, each in detail with a study in three parts; however, it is plausible that a number of limitations could have influenced the results obtained.

The user perception of the measures and their design alternatives was compared within-subjects in the first part of the study (section 4), but user-perceived usability evaluation was conducted between-subjects in the last part (section 6). Although this decision was justified due to the goals of the study, such as users evaluating the measures after ESM, yet it restricted the ability to thoroughly compare the usability of measures and their design alternatives without a baseline.

There is a possibility that participating in the second part of the study (section 5) after the first part (section 4) — in which users used all measurement tools within-subjects and expressed their preferences — and having several tasks in some cases impacted participants in unexpected ways, which were not investigated. For example, the gap between the two parts of the study varied for users. This gap depended on the exact start time of the participation for each user based on their choice and the time to the first scheduled sampling event. Participants specified their preference for the measurement tools. Then they were randomly assigned to use only one of the tools for the longitudinal measurements in the second part of the study. Therefore, if a participant, by chance, expected a different tool than the one they were randomly assigned to, still recalled their choice from the first part, and did not clearly read the study briefing and instructions, they could have ended up with an unfulfilled expectation. This disappointment could have turned into behaviors, such as early dropouts, etc., that were not investigated in this study.

In part II of this study (section 5), we attempted to consider the users' behavior in responding to a sampling event in the assessment of user compliance quantitatively through the weighted compliance variable. This enabled us to tax a user response that was submitted too quickly compared to their typical behavior in recording their mood. This strategy, in particular, was chosen since the focus of the samplings was mood measurement, in which users self-reflect about their internal states and then, accordingly, respond to the questions. This strategy might not necessarily be applicable to other types of experience sampling without further investigations and validation.

The qualitative assessment in the third part of the study (section 6) was based on the analysis of the written responses. Given that written responses are in general one-time feedback — compared to interviews — the obtained insights are limited to what participants explicitly stated. In other words, it was not possible to further explore the sentiment of the participants or assess the extent of their opinions and comments.

In this paper, we tested six mood measurement tools for smartphones as the medium. Only two of the most commonly used mood measures as representative of dimensional measures were considered, and the design alternatives were limited to only three designs per measure. Indeed, there are unlimited possibilities for the interaction design of a mood measurement tool and the medium involved. The wide choice of medium — e.g., use of wearables, mobile devices, artificial agents in the form of smart speakers or chatbots, robots, etc.—, various types of interaction methods — such as voice and gesture-based interaction—, and various techniques for increasing compliance in ESM — such as unlock-journaling, feedback presentation, etc.—, have added to the complexity and unlimited possibilities of the design space for mood measurement. This paper presented a comparable space between classic and newer tools by limiting the medium to smartphones and the conditions to the discussed measures and design alternatives using a smartphone app in a controlled and experimental setup. It also showed

the validity, compliance, and user experience dynamics and highlighted concerns arising from modifying the measurement tools. Nevertheless, the obtained results are bounded to, and impacted by, the tested conditions, and other designs, measures, mediums, and techniques should be further investigated. The characteristic of interaction is ultimately a choice that depends on the study design, yet it would not bring an exception for considering the assessment quality or app quality, which should be carefully discussed and examined.

8 CONCLUSION

This paper concentrated on mood tracking using smartphones and presented a study in three parts where the effects of mood measures (two chosen measures) and their design alternatives (classic, chatbot, and interactive designs) on (i) validity, (ii) user compliance, and (iii) user experience of a mood tracking app were investigated. The first part examined how classic mood measures and their design variations in the form of a chatbot and as an interactive design were compared with each other and how transferring and translating a mood measure impacted its obtained values, which can result in inaccurate results. The second part, using a two-week-long ESM study, explored the measures and their design alternatives in terms of user compliance defined by several variables. The final part investigated the user experience of those mood measures and design alternatives. As a result, we established that selecting and transferring a measure into a smartphone application is a process that requires careful consideration of both the user experience and usability of an app and its quality of assessments. We demonstrated that while one classic measurement tool is resilient toward modifications, another is susceptible to alterations. Therefore, a measure's resilience is an important factor to consider before using it. We also found that user compliance over time differed depending on the type of mood measure or its design; however, no design alternative was consistently better than others when both measures were considered. Our results further suggested that interactive design is more likely to attract users for a resilient measure and bring higher user compliance and user satisfaction. A measure, or its design, does not seem to significantly influence user perception of its user experience as long as the app is well-designed. In the end, choosing a measure or design is a trade-off between the measure accuracy, user compliance, and user satisfaction, which researchers have to prioritize. Ultimately, a successful mood measurement with smartphones has to consider both app quality and assessment quality and balance these two main facets.

ACKNOWLEDGMENTS

We would like to thank Prof. Dr. Jürgen Ziegler for his support. This paper is partially based upon research backed by the Interactive Systems Research Group.

REFERENCES

[1] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. Menthal: A Framework for Mobile Data Collection and Analysis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 624–629. https://doi.org/10.1145/2968219.2971591

[2] Daniel L. Ashbrook, James R. Clawson, Kent Lyons, Thad E. Starner, and Nirmal Patel. 2008. Quickdraw: the impact of mobility and on-body placement on device access time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '08)*. Association for Computing Machinery, 219–222. https://doi.org/10.1145/1357054.1357092

[3] 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

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating ScaleJUS. *Journal of Usability Studies* 4, 3 (May 2009), 114–123. https://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/

[5] Sonal Batra, Ross A. Baker, Tao Wang, Felicia Forma, Faith DiBiasi, and Timothy Peters-Strickland. 2017. Digital health technology for use in patients with serious mental illness: a systematic review of the literature. *Medical Devices (Auckland, N.Z.)* 10 (2017), 237–251. https://doi.org/10.2147/MDER.S144158

[6] Christopher J Beedie, Peter C Terry, and Andrew M Lane. 2005. Distinctions between emotion and mood. *COGNITION AND EMOTION* 19, 6 (2005), 847–878.

[7] Raymond Bond, Anne Moorhead, Maurice Mulvenna, Siobhan O'Neill, Courtney Potts, and Nuala Murphy. 2019. Behaviour Analytics of Users Completing Ecological Momentary Assessments in the Form of Mental Health Scales and Mood Logs on a Smartphone App. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (New York, NY, USA, 2019-09-10) *(ECCE 2019)*. Association for Computing Machinery, 203–206. https://doi.org/10.1145/3335082.3335111

[8] Mehdi Boukhechba, Lihua Cai, Philip I. Chow, Karl Fua, Matthew S. Gerber, Bethany A. Teachman, and Laura E. Barnes. 2018. Contextual Analysis to Understand Compliance with Smartphone-based Ecological Momentary Assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* (New York, NY, USA, 2018-05-21) *(PervasiveHealth '18)*. Association for Computing Machinery, 232–238. https://doi.org/10.1145/3240925.3240967

[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] Stuart K. Card, Thomas P. Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (July 1980), 396–410. https://doi.org/10.1145/358886.358895

[13] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D. Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (March 2018), 3:1–3:20. https://doi.org/10.1145/3191735

[14] 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

[15] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A. Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY, USA) *(UbiComp '15)*. Association for Computing Machinery, 121–132. https://doi.org/10.1145/2750858.2804266

[16] 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

[17] 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..

[18] M. Dubad, C. Winsper, C. Meyer, M. Livanou, and S. Marwaha. 2018. A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. *Psychological Medicine* 48, 2 (Jan. 2018), 208–228. https://doi.org/10.1017/S0033291717001659

[19] Gudrun Eisele, Hugo Vachon, Ginette Lafit, Peter Kuppens, Marlies Houben, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2022. The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. 29, 2 (2022), 136–151. https://doi.org/10.1177/1073191120957102 Publisher: SAGE Publications Inc.

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

[21] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[22] Karim El Batran and Mark D. Dunlop. 2014. Enhancing KLM (keystroke-level model) to fit touch screen mobile devices. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI '14)*. Association for Computing Machinery, Toronto, ON, Canada, 283–286. https://doi.org/10.1145/2628363.2628385

[23] Lisa Feldman Barrett and James A. Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology* 74, 4 (1998), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

[24] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT* 2 (2015), 6. https://doi.org/10.3389/fict.2015.00006

[25] Pascal E. Fortin, Yuxiang Huang, and Jeremy R. Cooperstock. 2019. Exploring the Use of Fingerprint Sensor Gestures for Unlock Journaling: A Comparison With Slide-to-X. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, Taipei, Taiwan, 1–8. https://doi.org/10.1145/3338286.3340135

[26] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. EMMA: An Emotion-Aware Wellbeing Chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. https://doi.org/10.1109/ACII.2019.8925455

ISSN: 2156-8111.

[27] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 8–14. https://doi.org/10.1109/ACII.2019.8925433 ISSN: 2156-8111.

[28] James J Gross. 2010. The future's so bright, I gotta wear shades. *Emotion Review* 2, 3 (2010), 212–216.

[29] 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

[30] 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

[31] Timothy J. Huelsman, J. R. Richard C. Nemanick, and David C. Munz. 1998. Scales to Measure Four Dimensions of Dispositional Mood: Positive Energy, Tiredness, Negative Activation, and Relaxation:. *Educational and Psychological Measurement* 58, 5 (Oct. 1998), 804–819. https://doi.org/10.1177/0013164498058005006 Publisher: Sage PublicationsSage CA: Thousand Oaks, CA.

[32] 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

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

[34] Rolf G. Jacob, Anne D. Simons, Stephen B. Manuck, Jeffrey M. Rohay, Shari Waldstein, and Constantine Gatsonis. 1989. The Circular Mood Scale: A new technique of measuring ambulatory mood. *Journal of Psychopathology and Behavioral Assessment* 11, 2 (1989), 153–173. https://doi.org/10.1007/BF00960477 Place: US Publisher: Plenum Publishing Corp..

[35] 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

[36] 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

[37] 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/.

[38] 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

[39] 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

[40] John D Mayer and Yvonne N Gaschke. 1988. The experience and meta-experience of mood. *Journal of personality and social psychology* 55, 1 (1988), 102.

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

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

[43] Margaret E. Morris, Qusai Kathawala, Todd K. Leen, Ethan E. Gorenstein, Farzin Guilak, William DeLeeuw, and Michael Labhard. 2010. Mobile Therapy: Case Study Evaluations of a Cell Phone Application for Emotional Self-Awareness. *Journal of Medical Internet Research* 12, 2 (2010), e10. https://doi.org/10.2196/jmir.1371

[44] 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

[45] 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

[46] 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

[47] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 92:1–92:16. https://doi.org/10.1145/3130957

[48] Aditya Ponnada, Binod Thapa-Chhetry, Justin Manjourides, and Stephen Intille. 2021. Measuring Criterion Validity of Microinteraction Ecological Momentary Assessment (Micro-EMA): Exploratory Pilot Study With Physical Activity Measurement. 9, 3 (2021), e23391. https://doi.org/10.2196/23391

[49] Andrew D. Rice and Jonathan W. Lartigue. 2014. Touch-level model (TLM): evolving KLM-GOMS for touchscreen and mobile devices. In *Proceedings of the 2014 ACM Southeast Regional Conference (ACM SE '14)*. Association for Computing Machinery, Kennesaw, Georgia, 1–6. https://doi.org/10.1145/2638404.2638532

[50] Aki Rintala, Martien Wampers, Ginette Lafit, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2021. Perceived disturbance and predictors thereof in studies using the experience sampling method. *Current Psychology* (June 2021). https://doi.org/10.1007/s12144-021-01974-3

[51] 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

[52] 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

[53] 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

[54] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[55] 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.

[56] Tijana Sagorac Gruichich, Juan Camilo David Gomez, Gabriel Zayas-Cabán, Melvin G. McInnis, and Amy L. Cochran. 2021. A digital self-report survey of mood for bipolar disorder. *Bipolar Disorders* 23, 8 (Dec. 2021), 810–820. https://doi.org/10.1111/bdi.13058

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

[58] 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

[59] 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

[60] Andreas Sonderegger, Klaus Heyden, Alain Chavaillaz, and Juergen Sauer. 2016. AniSAM & AniAvatar: Animated Visualizations of Affective States. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4828–4837. https://doi.org/10.1145/2858036.2858365

[61] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, Anchorage, AK, USA, 2886–2894. https://doi.org/10.1145/3292500.3330730

[62] 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

[63] 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

[64] Robert E Thayer. 1989. *The biopsychology of mood and arousal.* Oxford University Press.

[65] 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.

[66] Helma Torkamaan and Jürgen Ziegler. 2017. A taxonomy of mood research and its applications in computer science. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 421–426. https://doi.org/10.1109/ACII.2017.8273634 ISSN: 2156-8111.

[67] Helma Torkamaan and Jürgen Ziegler. 2020. Exploring chatbot user interfaces for mood measurement: a study of validity and user experience. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp-ISWC '20)*. Association for Computing Machinery, 135–138. https://doi.org/10.1145/3410530.3414395 1 citations (Crossref) [2022-04-05] event-place: New York, NY, USA.

[68] Helma Torkamaan and Jürgen Ziegler. 2020. Mobile Mood Tracking: An Investigation of Concise and Adaptive Measurement Instruments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 155:1–155:30. https://doi.org/10.1145/3432207

[69] Helma Torkamaan and Jürgen Ziegler. 2022. Recommendations as Challenges: Estimating Required Effort and User Ability for Health Behavior Change Recommendations. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM Press. https://doi.org/10.1145/3490099.3511118 0 citations (Crossref) [2022-04-05] event-place: Helsinki, Finland.

[70] A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, N. Palmius, M. Osipov, G. D. Clifford, G. M Goodwin, and M. De Vos. 2016. Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *Journal of Affective Disorders* 205 (Nov. 2016), 225–233. https://doi.org/10.1016/j.jad.2016.06.065

[71] Muhsin Ugur, Dvijesh Shastri, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Allison Kalpakci, Carla Sharp, and Ioannis Pavlidis. 2015. Evaluating smartphone-based user interface designs for a 2D psychological questionnaire. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, Osaka, Japan, 275–282. https://doi.org/10.1145/2750858.2805851

[72] 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

[73] Niels van Berkel and Vassilis Kostakos. 2021. Recommendations for Conducting Longitudinal Experience Sampling Studies. In *Advances in Longitudinal HCI Research*, Evangelos Karapanos, Jens Gerken, Jesper Kjeldskov, and Mikael B. Skov (Eds.). Springer International Publishing, 59–78. https://doi.org/10.1007/978-3-030-67322-2_4

[74] 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

[75] Chunpai Wang, Shaghayegh Sahebi, and Helma Torkamaan. 2021. STRETCH: Stress and Behavior Modeling with Tensor Decomposition of Heterogeneous Data. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21)*. ACM Press. https://doi.org/10.1145/10.1145/3486622.3493967 event-place: ESSENDON, VIC, Australia.

[76] 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

[77] David Watson. 2000. *Mood and Temperament*. Guilford Press. Google-Books-ID: iPFboulhcQcC.

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

[79] 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.

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

[81] 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://onlinelibrary.wiley.com/doi/abs/10.1002/9781118133880.hop202013

[82] Cheng K Fred Wen, Stefan Schneider, Arthur A Stone, and Donna Spruijt-Metz. 2017. Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* 19, 4 (April 2017). https://doi.org/10.2196/jmir.6641

[83] Marie T. Williams, Hayley Lewthwaite, François Fraysse, Alexandra Gajewska, Jordan Ignatavicius, and Katia Ferrar. 2021. Compliance With Mobile Ecological Momentary Assessment of Self-Reported Health-Related Behaviors and Psychological Constructs in Adults: Systematic Review and Meta-analysis. *Journal of Medical Internet Research* 23, 3 (March 2021), e17023. https://doi.org/10.2196/17023 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[84] C Wood and ME Magnello. 1992. Diurnal changes in perceptions of energy and mood. *Journal of the Royal Society of Medicine* 85, 4 (1992), 191–194.

[85] Xinghui Yan, Yuxuan Li, Bingjian Huang, Sun Young Park, and Mark W Newman. 2021. User Burden of Microinteractions: An In-lab Experiment Examining User Performance and Perceived Burden Related to In-situ Self-reporting. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (New York, NY, USA) *(MobileHCI '21)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3447526.3472046

[86] Xiaoyi Zhang, Laura R. Pina, and James Fogarty. 2016. Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness. 2016 (2016), 5658–5664. https://doi.org/10.1145/2858036.2858360

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