Exploring Chatbot User Interfaces for Mood Measurement: A Study of Validity and User Experience

Helma Torkamaan
University of Duisburg-Essen
Duisburg, Germany
h.torkamaan@acm.org

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

With the growth of interactive text or voice-enabled systems, such as intelligent personal assistants and chatbots, it is now possible to easily measure a user's mood using a conversation-based interaction instead of traditional questionnaires. However, it is still unclear if such mood measurements would be valid, akin to traditional measures, and user-engaging. Using smartphones, we compare in this paper two of the most popular traditional measures of mood: International PANAS-Short Form (I-PANAS-SF) and Affect Grid. For each of these measures, we then investigate the validity of mood measurement with a modified, chatbot-based user interface design. Our preliminary results suggest that some mood measures may not be resilient to modifications and that their alteration could lead to invalid, if not meaningless results. This exploratory paper then presents and discusses four voice-based mood tracker designs and summarizes user perception of and satisfaction with these tools.

KEYWORDS

Mood Tracking; Conversational ESM; PANAS; Affect Grid; Chatbot

1 INTRODUCTION

Mood tracking is of interest to various systems that need to model, understand, and predict human behavior, namely for the purpose of mental health, well-being, or building a mood-aware system. Current research-oriented mood-tracking apps try to accurately predict users mood and other behavioral phenomena, such as stress, using smartphone sensor data [12, 15]. However, to build such predictive models, one must rely mainly on accurate self-reported mood data that is captured mostly using experience sampling methods (ESM).

With the frequent use of smartphones as a platform for running ESM research, mood tracking using a smartphone has become a part of many tracking applications. Newer devices, such as wearables and both text and voice-enabled chatbots, have also shown potential for ESM and in-situ self-reporting. Although one can easily find several traditional valid measures of mood, these measures, mostly from the 1980s, were not originally designed to be used with smartphones or as conversational agents. When designing an ESM application (app) with various components, it may be essentially unfeasible to keep the original design of a measure — i.e. a questionnaire on paper. Researchers may choose to alter the design or presentation of a measure to attract users, make the app more interactive, prevent user fatigue, or utilize the full potential of smartphones. It is, however, unclear to what extent one can alter a measure and still have a valid assessment.

The validity of an altered measure is overlooked by some computer scientists whose concern is only the interactive design of an app or quick and effortless ESM. This can turn into an issue when an arbitrary mood measurement is used to build mood-aware systems or, especially, develop health-related solutions. Invalid mood measurements can lead to inaccurate or unjustified values of mood and consequently influence the overall outcomes or the predictive models. Therefore, exploring the validity and reliability of a modified measure, and the impacts of such alteration on the measurement and user experience, is a worthwhile endeavor.

2 RELATED WORK

Two of the most commonly used measures of mood are Affect Grid [11] and Positive and Negative Affect Schedule (PANAS) [14][1]. While both of these measures are based on the dimensional view on affect, Affect Grid is constructed on pleasantness-energy dimensions, whereas PANAS is based on the dimensions of positive activation (PA) and negative activation (NA). The dimensions used in both of these measures represent the same affective space, and PA-NA dimensions are only a 45° rotation of pleasantness-energy dimensions [11, 14]. It has been discovered that, despite this theoretical orientation of the dimensions, PANAS scores are not fully mappable to scores of Affect Grid [6, 11], particularly for scores of energy, in which the correlation of NA and energy is negligible. These two measures are commonly used for mood tracking, but it is hard to find a concurrent comparison between them on smartphones.

Russell et al. [11] indicated that Affect Grid could be used to capture both emotion and mood. The dimensions of

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

[1]According to Google Scholar, in June 2020 the citations count for Affect Grid is more than 1900, PANAS is more than 37,366, and I-PANAS-SF [13] is more than 1300.

pleasantness-energy are discussed predominantly in the affective computing community to describe and explore emotional states [2, 4, 9]. They also have been used for mood tracking, e.g., see [10, 12]. Due to Affect Grid's short format, capturing user responses is very quick and can be done with one touch on an app. However, the level of user training required for this measure, because of its complexity, can potentially lead to capturing inaccurate measurements. It should also be highlighted that if users can answer a measure with one touch, dismissing a sampling event would require the same — if not more — user effort as submitting a meaningless response, which might be the case for Affect Grid.

International PANAS short form (I-PANAS-SF) [13] is a shorter version of PANAS with ten questions. This measure is self-explanatory; however, it may not be the first choice for repeated assessments due to its longer length. Nevertheless, if participants submit a value with this measure, their responses are likely accurate because the effort required to provide a response would be much higher than that required to dismiss a sampling event. This measure has been used to capture users' mood using a smartphone, e.g., see [8].

The use of chatbots or conversational platforms, such as Alexa, to detect affect or to conduct ESM is a topic of interest for researchers [3, 5]. However, it is unclear if users prefer to interact with chatbots and conversation platforms, and if these methods capture valid assessments of mood. Accordingly, this paper first looks into the impacts of a text-based chatbot user interface (UI) on the validity of a measure in a smartphone-based mood-tracking app. The following section (§3) discusses the experiment design and preliminary results of this investigation, compares the two popular measures of mood — Affect Grid and I-PANAS-SF — on the app, and reveals how UI modifications impact the validity and user experience of a measure. Because chatbots can also be voice-enabled, we then focus on voice-based systems in order to fully examine the conversation-based UI. Section §4 explores the design and user satisfaction of four voice-based agents for mood tracking.

3 SMARTPHONE-BASED MOOD MEASURES

3.1 Method

Using a smartphone for mood tracking can alter the original mood measure. To find out whether a chatbot UI — instead of the original questionnaire form — influences the validity of a measure, we developed a mood-tracking app for Android phones. This app supports (figure 1) four mood measures: **(1)** I-PANAS-SF classic questionnaire; **(2)** I-PANAS-SF with a chatbot design; **(3)** Affect Grid classic grid view; and **(4)** Affect Grid with a chatbot design. When participants first use the app, they record their mood using all four of these measures. Recall (see §2) that I-PANAS-SF (measures 1 and 2) returns two values of PA and NA, whereas Affect Grid (measures 3 and 4) provides two scores total, one for pleasantness (PL) and one for energy (EN). The app then assigns each participant randomly to one of these four measures and, after two weeks of ESM, conducts a usability evaluation of each measure via a survey.

Our app has several considerations for the UI design of each measure. The classic versions of the measures are similar to the original pen-and-paper designs of the questionnaires. The chatbot UIs also try to stay loyal to the original design as much as possible by offering a uniform Likert-type response scale for each question. Unlike with pen-and-paper form, the app randomly orders the items (questions) of the I-PANAS-SF questionnaire [13] before each measurement. On another note, Affect Grid is a measure that originally has a graphical representation. To design its chatbot version (measure 4), we mainly relied on the report from Russell et al. [11] that indicated a strong correlation ($> .9$) between Affect Grid and separate single-item PL and EN questions. Accordingly, measure 4 uses this concept and asks separate questions for PL and EN.

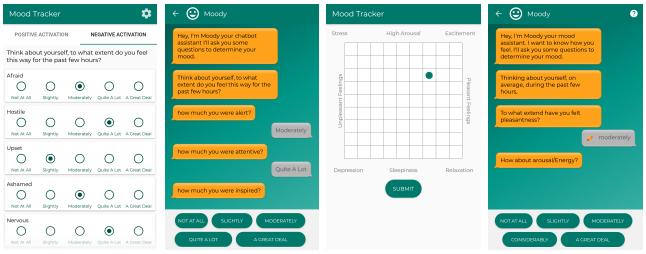

Figure 1: The app screenshots of the mood measures. Measures on the left are the I-PANAS-SF measures (1 and 2). Measures on the left are Affect Grid design measures (3 and 4)

In total, 425 participants (M: 121, F: 319) with an average age of $M = 32.42, SD = 11.32$ completed the preliminary within-subject evaluation of the four measures (randomly ordered). All measures ask participants about their mood in the past few hours. In total, 85 participants (M: 26, F: 58) finished the between-subject survey and expressed their opinions about the measures' usability.

3.2 Results and Discussion

We first compared the measures of I-PANAS-SF and Affect Grid in their classic design (measures 1 and 3) to ensure that, when used with a smartphone, they behave the same as in their original forms. The resulting score of PA weakly correlates with PL ($r = .49$, $p < .01$) and EN ($r = .39$). The score of NA has a correlation of $r = -.57$ with PL and $r = .25$ with EN. The observed correlation (table 1) was slightly stronger in our study with smartphones for PL and PA, NA and PL, and NA and EN, but had a slightly lower strength for PA and EN, compared to previous studies [7, 11]. This difference could be related to the use of I-PANAS-SF instead of PANAS in our case. It could also be an indication of a lower reliability of EN.

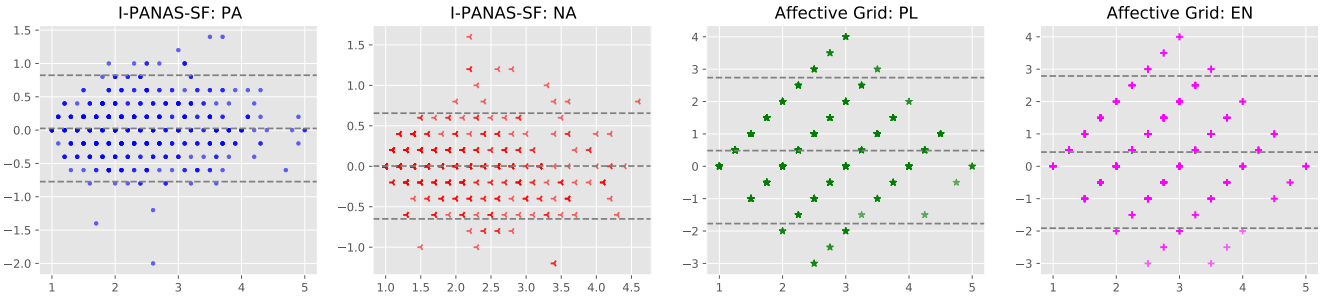

Figure 2: Distance between design alternatives for each value of measurement. PA: positive activation; NA: negative activation; PL: pleasantness; EN: energy; x-axis: average of two measures; y-axis: error or difference between measures

Table 1: The description of the primary variables across various measures and design alternatives of sample A.

	PA_1	NA_1	PA_2	NA_2	PL_3	EN_3	PL_4	EN_4
Mean	2.46	2.01	2.44	2.01	2.69	2.84	2.20	2.40
SD	0.89	0.82	0.87	0.85	1.16	1.12	1.04	1.05
Median	2.40	1.80	2.20	1.80	2.50	3.00	2.00	2.00
PA_1	-							
NA_1	-0.26	-						
PA_2	**0.89**	-0.21	-					
NA_2	-0.30	**0.92**	-0.23	-				
PL_3	0.49	-0.57	0.47	-0.59	-			
EN_3	0.39	0.25	0.38	0.23	-0.03	-		
PL_4	0.47	-0.32	0.46	-0.30	0.46	0.16	-	
EN_4	0.63	-0.17	0.63	-0.20	0.41	0.39	0.5	-

Note: PA: positive activation; NA: negative activation; PL: pleasure; EN: energy; 1: user groups 1: classic I-PANAS-SF; 2: chatbot I-PANAS-SF; 3: classic Affect Grid; 4: chatbot Affect Grid; all scores are scalled to be between 1–5.

As scores of I-PANAS-SF and Affect Grid are represent rotated dimensions, we used multiple regression to test their values and found that, despite their correlations, neither values of PA and NA nor PL and EN are strong predictors of each other. In fact, multiple correlation of Affect Grid with PA as a criterion is about .64, which is similar to the early findings of Russell et al. [11]. Nevertheless, we found a multiple correlation of .61 when NA was the criterion, which is much higher than what Russell et al. [11] reported. Therefore, in our sample, Affect Grid better predicts NA scores. The intercorrelation between dimensions of PL and EN is negligible in our findings. This suggests a high discriminant validity for Affect Grid — similar to [11]. PA and NA showed an intercorrelation of $-.23$, which is slightly lower than the reported intercorrelation in [13]. Altogether, these results suggest that the smartphone versions of these measures behave very similar to the original pen-and-paper versions and, therefore, can be reliably used with regards to one another.

The chatbot versions of these measures, however, suggest varying results. Measure 2 has a strong correlation (PA: $r = .89$ and NA: $r = .92$) with its classic counterpart, measure 1. The mean and median scores of PA and NA for these two measures are also very close (table 1). The difference between the two values of PA and NA from the classic (measure 1) and chatbot (measure 2) version of I-PANAS-SF — except for a few outliers — is consistently low (visualized in figure 2). In contrast, the scores of PL and EN from the classic Affect Grid (measure 3) have weak correlations with its chatbot (measure 4) version (PL: $r = .46$ and EN: $r = .39$). This correlation is even weaker than the associations of PL and EN with PA and NA. Measures 3 and 4 also have different mean and median values. As figure 2 shows, Affect Grid is generally less resilient than I-PANAS-SF, and in fact, the difference between its original and chatbot designs is more than expected. In short, altering Affect Grid may lead to invalid results. The results of this measure advise computer scientists against any alteration without validation, and encourage its careful use.

The usability evaluation of these measures did not reveal any major differences. The average SUS score [1] for all measures was consistently good. However, the classic version of Affect Grid (measure 3) had the lowest average score ($M = 76.14, SD = 19.12$), and I-PANAS-SF chatbot design had the highest ($M = 79.72, SD = 13.55$). The more detailed comparison showed that, compared to other measures, measure 3 has been considered better for frequent use, but worse for other variables of SUS. Measures 1 and 2 were considered better in their ease of use, lack of complexity, as well as need to learn. Surprisingly, users did not consider any of the measures to be very time-consuming; however, measure 4 was considered to be the least time-consuming. Future investigation of the qualitative responses could reveal more about these measures. Our initial assessment revealed no other differences between the measures. In other words, the design alternation of a measure did not significantly impact users' perception of its usability.

4 VOICE-BASED MOOD MEASURES

After examining the impact of a chatbot UI on the validity of a measure, we explored the design spaces of a voice-based mood tracking tool that can be used for ESM. After a thorough investigation of ESM and mood tracking, we ended up with a list of requirements for such a system. In short, the system should respect user privacy and, accordingly, allow users to access and review their personal data, and be activated

only with users' permission. The response format should either implement Likert-type scales or be based on open-ended questions. The system should provide a platform for receiving feedback from each user, give immediate feedback (or sympathy) upon users' responses, and support users with some sort of reminder, push notification, or routine. It should also use the voice-recognition functionalities of the platform to verify the identity of the participant. Researchers should be able to restrict and modify the timing and frequency of samplings. Finally, the system should start with a briefing, securely transfer user responses to the server, and end the final sampling with a debriefing.

We used these basic requirements to design and implement four working prototypes[2]: (A) I-PANAS-SF; (B) Affect Grid; (C) open-ended mood measure; and (D) hierarchical mood measure. The conversation starts with an initiation question, namely, *"Hello, got a minute for me?"* , *"Are you ready to tell me about how you are feeling?*, or *"Are you ready to answer my questions?"* etc. If a user declines a sampling event, the system response is a variation of this message: *"OK. Then please call me back when you are ready. I would like to know how you are feeling"*. The system uses several variations and synonyms for each sentence — which are not influential in the ESM results — to keep the conversation natural. If a user accepts the sampling event, the system continues with the measures' questions.

Of the four design alternatives, system A is basically the same as measure 2 in our Android app. The second system, B, is similar to measure 4 of our android app (see §3). The third system, C, is an open-ended measure that records users' mood states in their own words. To build this system, we collected a long list of mood states from the literature on affect and used it to represent each category of mood states or their synonyms. The main challenge for this approach was that many participants did not differentiate between mood and emotion, and they frequently used an emotional state to discuss their mood. Therefore, this system transfers and maps user responses about mood according to our predefined and trained list of adjectives for each mood state. The accuracy of system C can improve over time. The final system, D, is a hierarchical measure that assesses the valence of a users' mood by asking if they feel positive or negative. It asks about some specific mood states from PA and NA dimensions depending on the valence, and then ends the sampling by giving users the option to answer an open-ended question (a total of 4–8 questions). The goal of this prototype was to combine the strengths of the previous prototypes into one system.

Each of these systems has some advantages and drawbacks. For example, A can accurately capture user mood, its assessment is valid and comparable to PANAS, and it can be conducted in various languages. However, it is too long and can quickly cause user fatigue. In contrast, B is very short, but, as suggested by our smartphone-based study, it is not fully reflective of Affect Grid. Thus, it is inaccurate. Another short prototype is C, which could potentially attract users by letting them talk about their feelings. Unlike B, this system is unlikely to capture meaningless responses, since users would discuss their feelings and would not choose from a predefined answer set. However, mapping user responses to a valid measure with this measure and capturing the intensity of their mood states would be a challenge. This system does not benefit from showing users some mood states, which could help them to recall and describe their feelings better. Finally, D is shorter than A and has the advantages of both the open-ended questions and item-based measures.

To compare the user experience of these prototypes, we ran a small study with 16 university students as participants (M:7, F:9, average age: $M = 24.31, SD = 1.96$). Participants used all systems in random order and specified what they liked about each one. Qualitative analysis of the user responses indicates that, overall, users (n=11) liked the thoroughness of A and that it asks for specific affective states. A few users also suggested that, due to the level of detail in this measure, it would be sufficient for them to use it only once per day. Most users (n=13) liked the shortness of B, and reported (n=14) that the open-ended questions in their interaction with C felt natural and more pleasant. Almost half of the users (n=7) indicated that they liked the directions that D provided for the open-ended questions by also asking about specific mood states. During the experiment, 10 users indicated that they needed to relisten to the questions in A and B to fully understand them. In contrast, all participants clearly understood the questions in C and D. Most of the users (n=13) preferred to have a female voice for the agent, while the other 3 indicated that the voice's gender does not matter to them. Our experiment continued with several usability evaluations, e.g., assessing user-perceived accuracy, likeability, cognitive demands, annoyance, habituation, and speed. In the end, D seems to have the best usability and be the potential candidate for further investigation and validation.

5 CONCLUSION AND FUTURE WORK

This paper proposed a set of experiments to investigate the validity of mood measures using smartphones and conversational interactions. The result of a preliminary assessment (pre-study) with smartphone app suggests that I-PANAS-SF is a resilient measure toward UI modifications in the form of a chatbot, whereas Affect Grid is not. We then presented the requirements and four design alternatives of a voice-based mood tracker, and found which of those prototypes are better suited for validation with an ESM. In the future, we plan to compare our smartphone-based mood trackers with each other, as well as with our selected voice-based system in a longitudinal study; capture self-reported daily and overall weekly mood, as well as stress; and further investigate user compliance, validity, reliability, and predictive qualities of these measures.

[2]We used Google's Dialogflow platform to build each system's working prototype: https://cloud.google.com/dialogflow/docs

6 ACKNOWLEDGMENTS

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

REFERENCES

[1] John Brooke. 2013. SUS: A Retrospective. *J. Usability Studies* 8, 2 (Feb. 2013), 29–40. http://dl.acm.org/citation.cfm?id=2817912.2817913

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

[3] 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* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 58, 6 pages.

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

[5] 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)*. IEEE, 1–7.

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

[7] W. D. Killgore. 1998. The Affect Grid: a moderately valid, nonspecific measure of pleasure and arousal. *Psychological Reports* 83, 2 (Oct. 1998), 639–642.

[8] 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

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

[10] Verónica Rivera-Pelayo, Angela Fessl, Lars Müller, and Viktoria Pammer. 2017. Introducing Mood Self-Tracking at Work: Empirical Insights from Call Centers. *ACM Trans. Comput.-Hum. Interact.* 24, 1, Article 3 (Feb. 2017), 28 pages. https://doi.org/10.1145/3014058

[11] 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.

[12] 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

[13] 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.

[14] 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.

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