Evaluating Smartphone-based User Interface Designs for a 2D Psychological Questionnaire

Muhsin Ugur
Computational Physiology Lab
University of Houston
Houston, TX 77004 USA
mugur@uh.edu

Dvijesh Shastri
Dept. of Computer Science
and Engineering Technology
University of Houston - Downtown
Houston, TX 77002 USA
shastrid@uhd.edu

Panagiotis Tsiamyrtzis
Dept. of Statistics
Athens University of
Economics and Business
Athens, Greece
pt@aueb.gr

Malcolm Dcosta
Computational Physiology Lab
University of Houston
Houston, TX 77004 USA
mtdcosta2@uh.edu

Allison Kalpakci
Developmental
Psychopathology Lab
University of Houston
Houston, TX 77004 USA
ahkalpakci@uh.edu

Carla Sharp
Developmental
Psychopathology Lab
University of Houston
Houston, TX 77004 USA
csharp2@uh.edu

Ioannis Pavlidis
Computational Physiology Lab
University of Houston
Houston, TX 77004 USA
ipavlidis@uh.edu

ABSTRACT

This study explored various user interface designs to transition a two dimensional (2D) questionnaire from its paper-and-pencil testing format to the mobile platform. The current administration of the test limits its usage beyond the lab environment. Creating a mobile version would facilitate ubiquitous administration of the test. Yet, the mobile design must be at least as good as its paper-based counterpart in terms of input accuracy and user interaction efforts. We developed four user interface designs, each of which featured a specific interaction approach. These approaches included displaying the 2D space of the questionnaire in its original form (M1), inputting one variable at a time on the 2D space (M2), dissolving the 2D space into two one-dimensional ordinal scales (M3), and orienting the input selections to the diagonal axes (M4). The designs were tested by a total of 34 participants, aged 18 to 52 years. The study results find the first three interaction approaches (M1-M3) effective but the fourth approach inefficient. Furthermore, the results indicate that the two-tap designs (M2 and M3) are equally as good as the one-tap design (M1).

UbiComp '15, September 7–11, 2015, Osaka, Japan.

http://dx.doi.org/10.1145/2750858.2805851

Author Keywords

Online questionnaires; 2D questionnaires; Mobile devices; Mobile health care; User interface design.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation: User Interfaces - Evaluation/methodology, Input devices and strategies, Prototyping, Interaction styles.

INTRODUCTION

Questionnaires are important elements of psychological and social science studies. Traditionally, questionnaires are paper-and-pencil tests where study participants are required to mark their responses on paper. Recent proliferation of online survey tools (e.g., SurveyMonkey, Qualtrics, and Google consumer surveys) has led to the replacement of the paper-based administration of the tests with online testing. The tools offer the convenience of completing these questionnaires from anywhere at any time and on various computing platforms including desktop, laptop, and mobile. Thus, they facilitate ubiquitous administration of these tests. The online tools are also convenient for the study experimenters as they provide the experimenters with instant access to the participants' data. The data can be instantly logged to online servers

and immediately made available to the experimenters for data analysis.

Some of these questionnaires aim to collect factual information. A demographic questionnaire is an example of this type. A typical demographic questionnaire collects information about a person's age, gender, education level, race, etc. Responses to these types of measures typically do not fluctuate with time, and hence, data input errors are relatively easy to rectify for these questionnaires.

In addition, psychological and social science studies assess people's traits and states including emotions, attitudes, and perceptions for certain entities. These questionnaires are referred to as psychological questionnaires. The Positive and Negative Affect Schedule (PANAS), and the State-Trait Anxiety Inventory (STAI) are well-known examples. Responses to variables on these questionnaires may be time bound. Therefore, it is important to collect responses temporally close to their actual occurrence. Equally important is to design an online user interface that allows the participants to input their responses as accurately as possible because the responses to the psychological variables may be vulnerable to retrospective recall bias [3].

A fair number of studies have been conducted to test the validity and reliability of online questionnaires and the usability of their user interface designs to facilitate factual information collection [2]. However, lesser attention has been paid to online versions of psychological questionnaires.

Carlbring *et al.* investigated the effectiveness of online versions of a panic disorder measure [1]. Zhu *et al.* evaluated the effectiveness of a mobile-based questionnaire designed to identify children suffering from post-traumatic stress disorder (PTSD) [10]. Shaik, Wong and Teo in their recent work reported how the layout of psychological questionnaires and national culture affect people's responses to online questionnaires [8]. Väätäjä and Virpi discussed user interface design guidelines specific to smartphone-based questionnaires [6]. The guidelines were tested on a questionnaire that aimed to report users' emotions and satisfaction while interacting with a mobile journalism application.

The common denominator of these studies is the structure of their questionnaires. Typically, these questionnaires are multi-question questionnaires in which the users are allowed to enter their responses for each question one at a time. Essentially, this is a one dimensional (1D) design because every study variable is treated independently. Typically, an ordinal scale is used to record the level of agreement or disagreement for each study variable. The questionnaires can be either textually (e.g., [1], [8], [10]) or visually (e.g., [6]) represented. The primary user interface (UI) design challenge is the layout of the questionnaire to optimize input speed and accuracy.

Our study focuses on UI design optimization for a two dimensional (2D) questionnaire. Two dimensional questionnaires are common in psychological studies. Typically, a 2D questionnaire is formed by simultaneously presenting two (or more) study variables on a 2D space. The space is discretized into a 2D grid in which each axis represents one study variable. Usually, it is presented in a paper-and-pencil format. Participants are required to record their responses by marking a single cell on the grid.

In 2010, Morris *et. al* presented a mobile version of a 2D questionnaire [4]. Specifically, they developed a mobile application called Mood Map which allows the users to record their moods several times a day. Mood Map features a 2D grid formed by the horizontal axis representing *mood* dimension and the vertical axis representing *energy* dimension. Their study aimed to explore the feasibility of delivering psychotherapies via the mobile platform. Little attention, however, was given to the designing of the grid interface.

In this paper, we propose four different UI designs for a 2D questionnaire. The questionnaire assesses participants' perceptions about others' interpersonal behavior. Through these designs, we explore various ways to interact with a 2D questionnaire on the mobile platform. The fundamental designing question is *how to optimize the user interface of the grid so that the users can record their inputs as accurately as possible with minimum input efforts.*

In the remaining paper, we first introduce the 2D questionnaire for which we developed four mobile interfaces. Next, we discuss the experimental design, the four user interface designs, and study results. Finally, we conclude the paper with the research summary.

THE 2D PSYCHOLOGICAL QUESTIONNAIRE

The 2D questionnaire that we used in the study is a well-known measure for assessing interpersonal perceptions based on the interpersonal circumplex [9]. The interpersonal circumplex organizes interpersonal dispositions in a 2D space, with a communal dimension along the horizontal axis and an agentic dimension along the vertical axis [5]. Figure 1 shows the layout of the grid. The grid is arranged in 11 columns and 11 rows depicting 11 discrete levels of the communal and agentic dimensions, respectively. The communal dimension represents efforts that augment affiliation and interpersonal connectedness. It is ranged from *Cold-Quarrelsome* interaction on the left to *Warm-Agreeable* interaction on the right of the grid. The agentic dimension represents efforts that serve desires for autonomy and control. It is ranged from *Assured-Dominant* interaction on the top to *Unassured-Submissive* interaction on the bottom of the grid.

Traditionally, the interpersonal grid has been used in a paper-and-pencil format. Participants describe their perceptions of others' behavior by marking their responses in a single cell of the grid, indicating the extent to which the other person was perceived as agreeable (vs. quarrelsome) and as dominant (vs. submissive) in a specific interaction. The grid features four possible interaction outcomes: 1) *Engaging interaction* anchored at the top-right corner of the grid, 2) *Critical interaction* anchored at the top-left corner of the grid, 3) *Withdrawn interaction* anchored at the bottom-left corner of the grid, and 4) *Deferring interaction* anchored at the bottom-right corner of the grid. Reliability and validity of the Interpersonal Grid is presented in [5].

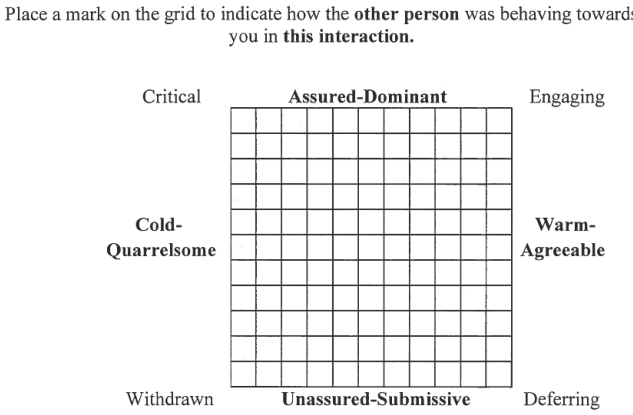


Figure 1: The 2D psychological questionnaire.

(a) Paper trial (b) Mobile trial

Figure 2: Experimental setup

EXPERIMENTAL DESIGN

Participants

A total of $n = 34$ participants (17 males + 17 females) volunteered for this study. Their ages ranged between 18 and 52 years ($\mu \pm \sigma = 30.5 \pm 9.4$). The participants were recruited within the University of Houston's premises via emails and advertisement flyers. They received a $20 gift-card for their participation. The study was approved by the university's institutional review board.

Procedure

The study participants paid a single visit to the experiment. At the beginning of the experiment, participants were asked to complete the study consent form, a demographic form, and a smartphone familiarity form. About a minute after that, the experiment began. The experiment was divided into five trials - one trial for the paper-and-pencil test, and the remaining four trials for the mobile versions. The paper version was treated as the gold-standard, against which the performance of the four mobile designs were compared. The order of the trials was randomized in a Latin Square fashion to counterbalance the order effect. Two successive trials were separated by about a minute long break in which the participants were asked to rest.

Figure 2 (a) shows the experimental setup for the paper trial, and Figure 2 (b) shows the experimental setup for the mobile trials. In each of these trials the participants were asked to report their most recent social interaction before arriving for the experiment. The interpersonal circumplex defines a social interaction as an interaction with one other individual lasting five or more minutes. By allowing the interactions to be outside the experimental setup, we were able to evaluate a majority of the spatial area of the grid.

Before completing the grid, the participants completed two other questionnaires for all five trials. The first questionnaire collected the participant's current emotional states on a seven-point scale. Specifically, the questionnaire recorded 12 emotions including happy, pleased, worried, depressed, ashamed, etc. The second questionnaire collected the information of the individual with whom the participant interacted. Specifically, it recorded the individual's age, gender, and the relationship with the participant. These two questionnaires were laid out in the exact same fashion in all four mobile designs. Therefore, they were not included in the design analysis. They were part of the experiment primarily to avoid having the participants carelessly input their responses for the grid. The delivery order of the three questionnaires was the same for all five trials.

In the paper trial, the participants were asked to complete these three questionnaires in the paper-and-pencil format (See Figure 2 (a)). Each questionnaire was printed on a separate 8-by-11 inch paper. The time spent on each questionnaire was recorded.

In the mobile trials, the participants were asked to complete the same three questionnaires on an iPhone 5. The 2D questionnaire is typically completed six times a day in clinical studies. Therefore, it is more practical to use a smartphone than a tablet for the mobile versions. To facilitate the mobile trials, we developed an app in iOS 8. The app features three UI views, one per questionnaire. The views were presented in the same order as the paper trial. Moreover, the first two views did not differ between the mobile trials as they were not the focus of the UI design. The third view presented the 2D grid. We developed four UI designs for the 2D grid. Each of the four mobile trials presented one of the four grid designs. The responses to these grid designs were later evaluated against the paper version.

For each UI design, the app logged three values: time spent on the grid, number of taps, and final cell selection. The time spent is defined as elapsed time between entering the grid view screen and proceeding to the next screen. The number of taps is defined as the total number of taps that the user made on the grid for inputing his/her response.

The participants were required to hold the phone in the portrait fashion, as this orientation matched with its paper-based counterpart. Before the beginning of the first mobile trial, the

experimenter instructed the participants on how to proceed between the multiple views of the app.

At the end of each trial, the participants were asked to complete the NASA task load index (NASA-TLX) questionnaire and a usability questionnaire. The NASA task load index evaluated the participants' perceptions about their interaction efforts with the designs. The usability questionnaire evaluated the UI designs from the visual aesthetic and operating convenience standpoints.

USER INTERFACE DESIGNS

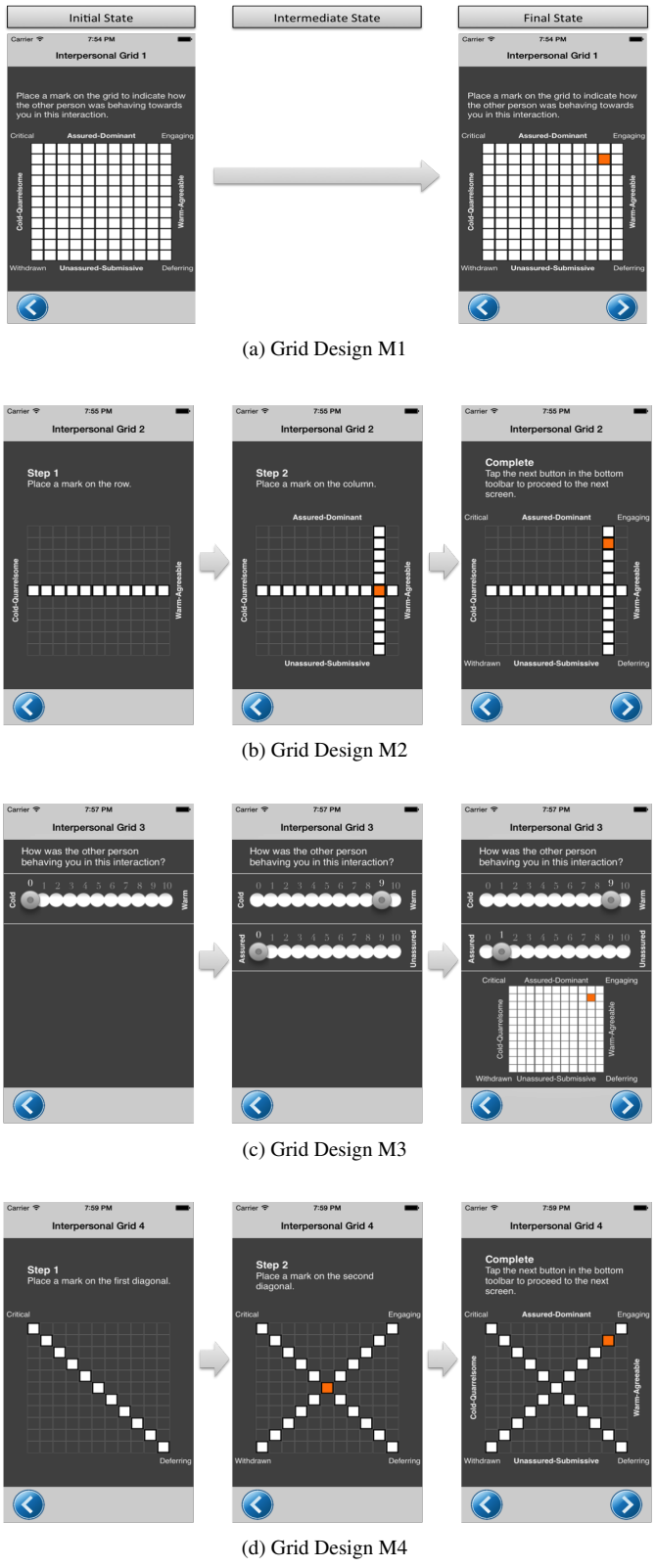
Figure 3 illustrates the four UI designs of the 2D questionnaire. The first design (M1) is a replica of the 2D questionnaire's paper-and-pencil format (see Figure 3 (a)). The users had to tap at least once to input their responses. Thus, M1 features a one-tap design. Refining or reentering of the response requires the grid to be reset first by deselecting its current selection.

The remaining three designs are two-tap designs. The users had to tap at least twice to input their responses. By adding one additional tap, we aimed to achieve improved usability without sacrificing input accuracy. The two-tap designs ease the cognitive demand by allowing the users to input their responses for one variable at a time.

The second design (M2) preserves the spatial visualization of the grid (see Figure 3 (b)). It demands two taps in series: the first tap is to input a response for the communal dimension, and the second tap is to input a response for the agentic dimension. The interface guides the users by first enabling the cells in the middle row (see the left view), and then enabling the cells in the corresponding column (see the middle view). The input process is illustrated in Figure 3 (b). Refining or reentering of the response requires the grid to be reset to its initial state. This happens in the reverse order in which the user has to first deselect the current row selection (the agentic dimension) and then the current column selection (the communal dimension).

The third design (M3) dissolves the spatial representation of the grid (see Figure 3 (c)). It presents the 2D grid as two 1D ordinal scales, one scale per grid dimension. This representation is common for online questionnaires, and hence the user would have familiarity of its handling. A slider object control of iOS 8 is used as an input interface. The design demands two taps in series: the first tap is to input a communal response and the second tap is to input an agentic response. The final selection is superimposed on the grid. The input process is illustrated in Figure 3 (c). The response refinement or reselection can be achieved in any order via the sliders.

The fourth design (M4) is similar to the second design with only difference is that it features two diagonal selections (see Figure 3 (d)). Here, the users can directly input their responses for the interaction outcomes. Since the questionnaire deduces a social interaction into critical, deferring, withdrawn, or engaging interactions, we decided to let the users input their responses for these outcomes directly. The design demands two taps in series: the first tap is to input the



(a) Grid Design M1

(b) Grid Design M2

(c) Grid Design M3

(d) Grid Design M4

Figure 3: Four UI designs of the 2D questionnaire.

response for the *Critical - Deferring* interactions, and the sec-

ond tap is to input the response for the *Withdrawn - Engaging* interactions. The input process is illustrated in Figure 3 (d). The response refinement is achieved in the reverse order similar to M2.

APP ARCHITECTURE

We developed an iPhone app to support the four UI designs. Figure 4 shows a schematic diagram of the app framework. The framework features the client-server architecture. The web server uses PHP scripts for handling the client's requests and MySQL database for data storage. The database consists of two tables: User table and Log table. The user table stores the user's ID, whereas the log table stores the user interaction data including grid ID, time spent on grid, number of taps, and the final cell selection.

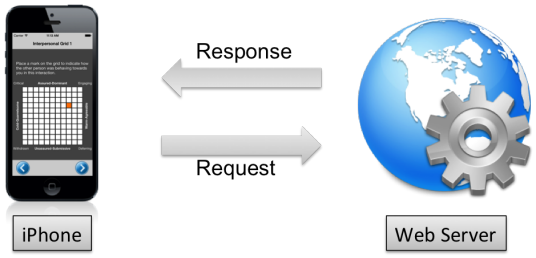


Figure 4: App Architecture

Once the users input their data, the app passes the data along with the other necessary parameters to the web server. The web server accepts data from the iPhone client, logs the data, and sends back appropriate messages to the iPhone client in the JSON format.

DATA ANALYSIS

When factual information is collected, the aim is to collect responses that are deliberate and accurate, but in case of the psychological information the aim is to collect responses that are spontaneous [7]. Unlike factual information, spontaneous responses are vulnerable to retrospective recall bias. Furthermore, mobile questionnaires are meant for data collection outside the lab environment where the experimenters are not available to rectify the users' mistakes. Therefore, a user interface for a psychological questionnaire should be as intuitive as possible to minimize the design ambiguity and maximize input accuracy. Moreover, the interface should not be perceived as mentally demanding. Otherwise, the users may not be able to provide in-the-moment responses consistently by delaying the interaction or avoiding the interaction completely. Both actions are detrimental to the study because the former action potentially logs inaccurate responses while the latter action generates missing values. Considering these factors, we evaluated the proposed UI designs along the three axes: User input consistency, user input effort, and user perception.

User Input Consistency

An ambiguous user interface may cause the users to input their responses differently from what they would enter for the paper-based test. Thus, by comparing the users' mobile inputs against their inputs for the paper-based test, one can identify design ambiguity. Figure 5 illustrates the grid inputs for the entire study population. Each grid shows a participant's five responses - one response for the paper trial and four responses for the mobile trials. The figure visualizes all the input responses in a single view. Two observations can be drawn from it. First, the input responses are all over the grid indicating that the study included a majority of the interaction outcomes. Second, except for one case (P6), the input responses vary by design.

We performed statistical analysis to test the significance of the response variation. Assuming the paper-based test as ground truth, we computed Manhattan distance between each mobile design input and paper input. The boxplots in Figure 6 present these distances for the entire study population. The distance measurement signifies input consistency of a mobile design. A lower distance value indicates lower input error and hence, higher input consistency, whereas a higher distance value indicates lower input consistency. Among the four designs, M1 shows the highest input consistency while M4 shows the lowest input consistency.

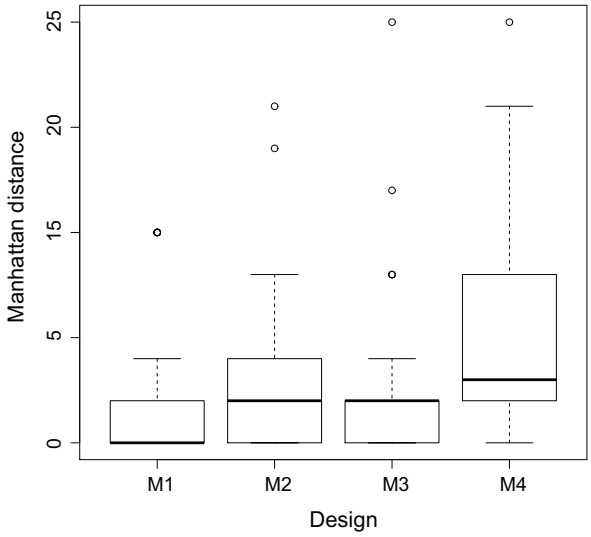


Figure 6: Boxplot diagrams represent the Manhattan distance between the paper inputs and each mobile design inputs for the entire study population ($n = 34$).

We performed repeated-measures ANOVA test on the distance variable. The test reveals significant mean differences in input consistency among the four designs (mixed effects ANOVA, $p = 0.0010$). Post analysis indicates that M4 has significantly higher mean Manhattan distance error than the rest of the designs, indicating that M4 has the lowest input consistency among the proposed designs. Remaining three designs (M1-M3) do not have significant mean Manhattan error differences (mixed effects ANOVA, $p = 0.3926$).

User Input Effort

We computed the number of excessive taps that the participants made before finalizing their responses. Specifically, we

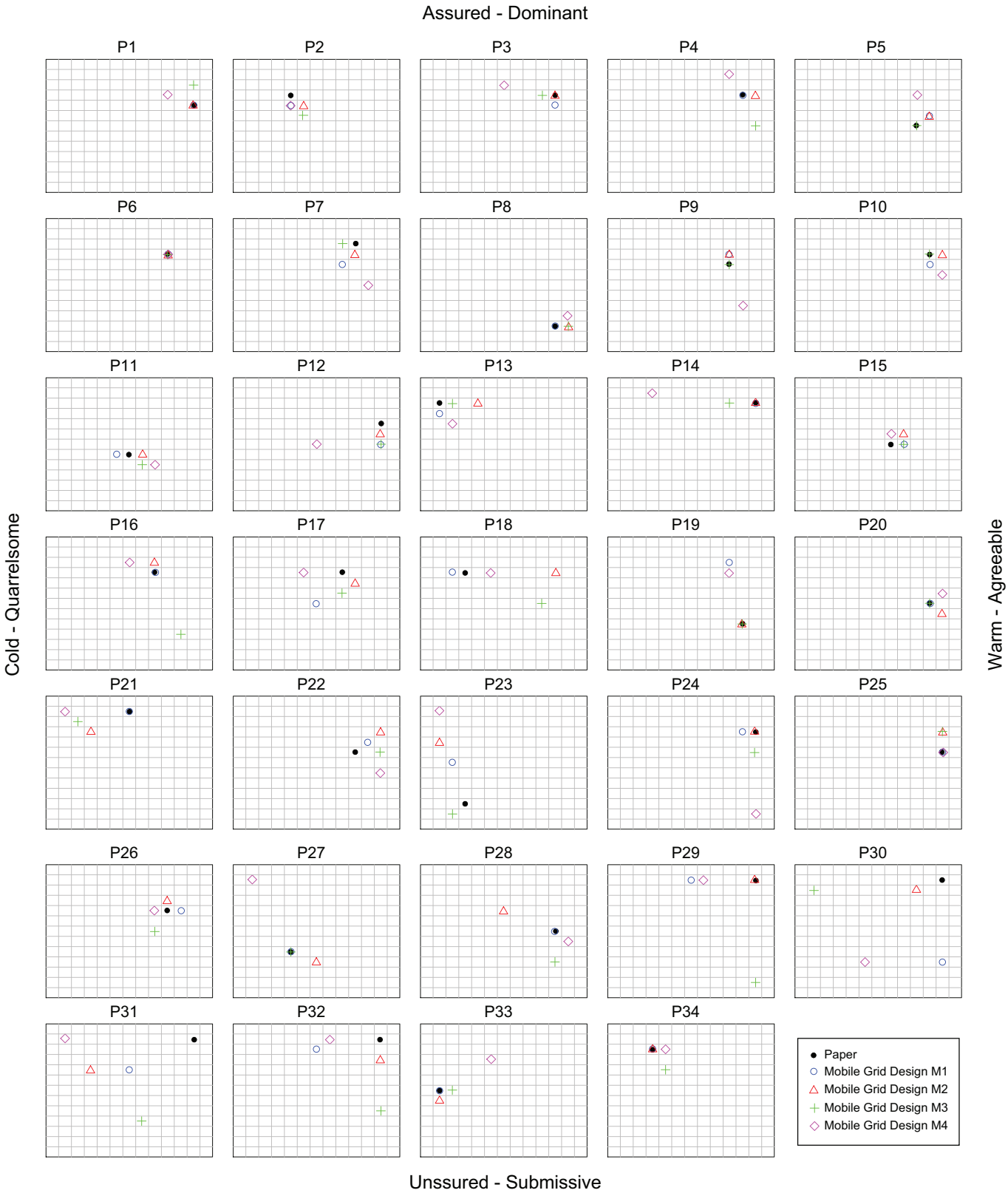


Figure 5: The grid inputs of the entire study population ($n = 34$). Each grid shows a participant's five responses including one response for the paper version and four responses for the mobile versions (M1-M4).

subtracted the minimum required taps per design from the total number of taps made per design by each participant. Higher number of excessive steps indicates more input effort from the user. We ran repeated-measures ANOVA test on all four designs. The test shows that the variability in the excessive number of taps for all four designs is statistically insignificant ($p = 0.3802$). Thus, all the designs demand on average the same amount of effort from the users.

User Perception

Analysis of the NASA-TLX questionnaire:
To evaluate how the participants perceived their interactions with each design, we analyzed their responses to the NASA-TLX questionnaire. The questionnaire evaluated their perceptions for six variables: mental demand, physical demand, temporal demand, performance, effort, and frustration.

We performed repeated-measures ANOVA tests for each variable. The analysis reveals that the participants perceived the M4 design significantly mentally demanding ($p = 0.0262$). We believe that the design ambiguity of M4 may have made the interaction mentally demanding for the participants. The other five variables, physical demand ($p = 0.1014$), temporal demand ($p = 0.8138$), performance ($p = 0.2775$), effort ($p = 0.1148$), and frustration ($p = 0.1477$) were not statistically different among all four designs. The boxplots in Figure 7 present NASA-TLX scores for the entire study population.

Analysis of the usability questionnaire:
To evaluate how the participants perceived the usability of each design, we analyzed their responses to the usability questionnaire. The questionnaire included the following questions:

Q1. The interface is aesthetic. Agree?
Q2. The interface is ambiguous. It took me a while to understand how to input my response. Agree?
Q3. The interface is easy to use. Agree?
Q4. I learned to use it quickly. Agree?

We performed non-parametric Kruskal Wallis test for each variable. The test reveals that the distributions of responses to Q1 for all four designs are not statistically different ($p = 0.1225$), which indicates that the aesthetics of M1-M4 do not differ significantly. The test for Q2, Q3 and Q4 reports significant difference among the 4 designs ($p = 0.0016$ for Q2, $p = 0.0014$ for Q3, and $p = 0.0013$ for Q4). Excluding M4 in the follow-up analysis reveals that the distributions of responses to Q2, Q3 and Q4 for designs M1-M3 are not statistically different ($p = 0.624$ for Q2, $p = 0.2686$ for Q3, and $p = 0.709$ for Q4). These indicate that M4 was perceived ambiguous, difficult to use and hard to learn. The bar charts in Figure 8 present usability scores for the entire study population.

CONCLUSION

This research explores various ways to interact with a 2D questionnaire on the mobile platform. Two dimensional questionnaires are common in psychological studies. Yet, most questionnaires are administered in paper-and-pencil format.

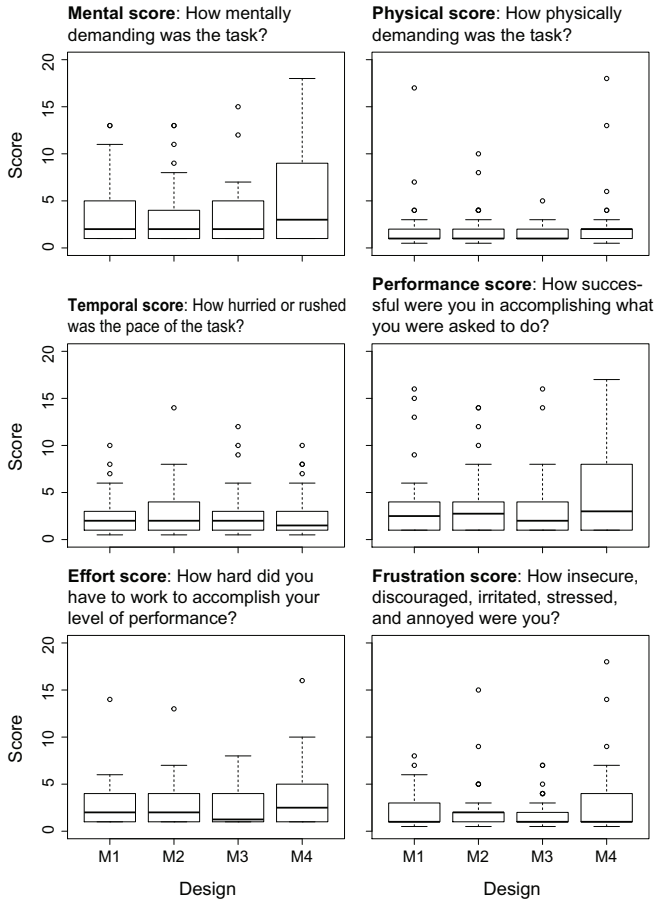


Figure 7: Boxplot diagrams represent analysis of the NASA Task Load Index questionnaire for the entire study population ($n = 34$). The participants were required to respond to each of these variables on a 20-point scale where 1 being *very low* and 20 being *very high*.

This is an inconvenient approach, particularly when a questionnaire needs to be completed multiple times a day. Having the questionnaire on the mobile platform allows ubiquitous data collection. The challenge, however, is to design intuitive user interfaces that facilitate seamless user interactions. This requires consideration of the smartphone's viewing space, which is much smaller than the standard paper-and-pencil testing format.

In addition, the UI design should take into account the interaction efforts demanded by a 2D questionnaire. Typically, a 2D questionnaire is formed by simultaneously presenting two (or more) study variables on a 2D space. The users have to input the variables' values in a single selection. This is fundamentally different from most online questionnaires in which each study variable is treated independently, and presented on a 1D ordinal scale. Handling a 2D questionnaire is cognitively more demanding.

The proposed UI designs are developed with having these considerations in mind. Specifically, we developed four UI designs for a 2D psychological questionnaire which is a well-

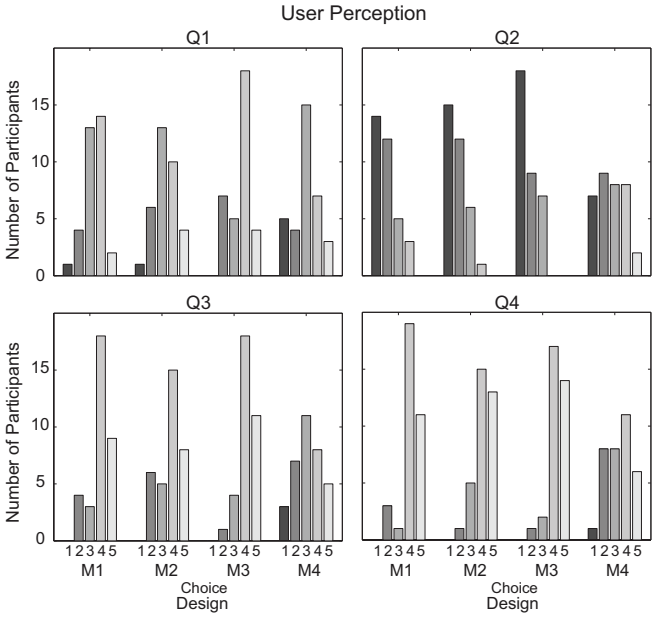


Figure 8: Bar graphs represent analysis of the usability questionnaire for the entire study population ($n = 34$). The participants responded to each of these questions on a five-point scale where 1 being *Not at all* and 5 being *completely*.

known measure for assessing participants' perceptions about others' interpersonal behavior. Each design featured a specific interaction approach. Design M1 is a replica of the 2D questionnaire's paper test. Design M2 preserves the spatial visualization of the questionnaire but lets the user handle one variable at a time. Design M3 dissolves the spatial visualization and presents both variables on two 1D ordinal scales. This representation is common for the online questionnaire and hence, the users would find it easy to interact with. The last design, M4, orients M2 for diagonal selections. In place of the two orthogonal dimensions (communal and agentic), this design allows selections of their interactions (engaging, critical, withdrawn, and deferring) directly. All the designs were tested by a total of 34 participants.

The study results show that M4 is the most inefficient user interface design. The participants not only made more mistakes while entering their responses but also perceived the design mentally demanding, ambiguous, difficult to learn and use. Two lessons can be learned from this finding. First, a user interface design featuring meta-data input should be discouraged. Second, inputing responses along the diagonal dimensions is not a conventional practice and hence, should be used with caution. A combination of both could be an ineffective UI design such as M4.

The performance across the remaining three designs (M1-M3) is not statistically different indicating that a two-tap design (M2-M3) was not necessarily better than a one-tap design (M1). A field study in which participations are required to input their responses multiple times a day is necessary to further verify this finding.

In the near future, we plan on expanding these designs to a field study in which the users will input their responses six times a day for an extended period of time. The field study will not include the M4 design, since it has been observed as the most inefficient user interface design.

ACKNOWLEDGMENTS

This research was supported by the Eckhard-Pfeiffer Distinguished Professorship fund of Ioannis Pavlidis. We would like to thank all the volunteers who participated in our study.

REFERENCES

1. Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., and Andersson, G. Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior 23*, 3 (2007), 1421–1434.
2. Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American Psychologist 59*, 2 (2004), 93.
3. Hufford, M. R., Shiffman, S., Paty, J., and Stone, A. A. Ecological momentary assessment: Real-world, real-time measurement of patient experience.
4. Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., and Deleeuw, W. Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. *Journal of medical Internet research 12*, 2 (2010).
5. Moskowitz, D., and Zuroff, D. C. Assessing interpersonal perceptions using the interpersonal grid. *Psychological assessment 17*, 2 (2005), 218.
6. Väätäjä, H., and Roto, V. Mobile questionnaires for user experience evaluation. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, ACM (2010), 3361–3366.
7. Van Schaik, P., and Ling, J. Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction (TOCHI) 14*, 1 (2007), 4.
8. Van Schaik, P., Wong, S. L., and Teo, T. Questionnaire layout and national culture in online psychometrics. *International Journal of Human-Computer Studies 73* (2015), 52–65.
9. Wiggins, J. S. Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior.
10. Zhu, Z.-H., Huang, F., Wang, W.-Z., Zhang, J.-X., Ji, Y., and Zhang, K. The psychometric properties of children's impact of event scale administered via mobile phone. In *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on*, IEEE (2009), 1–3.