Inviting Participants' Peers in a Mobile Assessment Study: An Empirical Investigation

Yu-Lin Chang
lingo1995@gapp.nthu.edu.tw
Department of Computer Science, National Tsing Hua University
Hsinchu, Taiwan

Hao-Ping (Hank) Lee
hankhplee@gmail.com
School of Interactive Computing, Georgia Institute of Technology
Atlanta, GA, USA
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, Taiwan

Yung-Ju Chang
armuro@cs.nctu.edu.tw
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, Taiwan

Chih-Ya Shen
chihya@cs.nthu.edu.tw
Department of Computer Science, National Tsing Hua University
Hsinchu, Taiwan

ABSTRACT

Mobile assessment is commonly adopted to obtain information about individuals' statuses, but is limited by the participants' receptivity to assessment prompts. This study explores the feasibility of participants in such studies recruiting their peers to help report their locations, activities, and emotions. Over a two-week period, 15 main participants and a total of 82 of their peers collaboratively provided mobile assessments. We showed that when the main participants were not receptive to assessment prompts, their peers provided the requested information in 96% of cases, with 42% of the time feeling confident in their assessments. However, the peers' levels of confidence and agreement with one another both varied by assessment-question type. Location information was provided the most confidently, but the latter was most likely to agree with the participants' own assessment. We also discuss matrices, including of agreement rate and peer numbers, that future peer-assisted mobile-assessment research should consider.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Mobile Assessment; Data Quantity; Data Quality

ACM Reference Format:
Yu-Lin Chang, Hao-Ping (Hank) Lee, Yung-Ju Chang, and Chih-Ya Shen. 2021. Inviting Participants' Peers in a Mobile Assessment Study: An Empirical Investigation. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21), September 27-October 1, 2021, Toulouse & Virtual, France.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3447526.3472021

MobileHCI '21, September 27-October 1, 2021, Toulouse & Virtual, France

ACM ISBN 978-1-4503-8328-8/21/09…$15.00
https://doi.org/10.1145/3447526.3472021

1 INTRODUCTION

Mobile assessment (MA) is widely used in the field of human-computer interaction to obtain status or context information about study participants. While some researchers have leveraged it to obtain users' in-situ experiences and contextual information at specific moments – an approach commonly referred to as the experience sampling method (ESM) [10, 38] or ecological momentary assessment (EMA) [35] – others have used it to label participants' location, activity, or emotional state [6, 7, 32], which is referred to as mobile prompted labeling. One prominent use of MA has been to inform the building of predictive models [18].

To ensure that MA responses are truthful and not subject to serious recall bias/error, however, such approaches generally require that their assessment prompts be responded to within a specified, fairly short period, and responses received beyond that time limit are deemed invalid [28]. This obviates a problem that is all but unavoidable in other retrospective methods such as diary studies and interviews [28]. By their nature, however, such time limits rely on participants being receptive to prompts more or less at the moment they receive them [25]. This potentially introduces another type of bias into the responses obtained: toward moments of high receptivity to being interrupted by mobile prompts [15]. Capturing more data from low-receptivity moments by extending the number of days of data collection might seem like a straightforward solution, but it is not ideal, because the sheer number of assessment prompts issued to each participant – often up to 12 per day – is burdensome [20, 42] and can even result in their compliance with prompts decaying gradually over time [34]. This dynamic may negatively impact not only the quantity, but also the truthfulness, of their responses [34, 41].

Researchers have therefore explored several other approaches to enhancing MA participants' response rates, such as visualization-based feedback [17], gamification [43], and changes to data entry [39, 47, 48]. Some studies have used mobile-phone sensor data to augment user-provided information [4, 42] or to detect opportune moments for triggering assessment prompts [26, 44]. However, none have solved the fundamental receptivity issue noted above: that MA data can only be obtained if and when participants are available to respond to an assessment questionnaire. Accordingly, the current study proceeds from the insight that when a participant is unavailable, it might be feasible to allow selected other people who know something about the main participant's situation, hereafter referred to as peers (cf. [3]), to answer certain questions on his/her behalf, thus increasing the chances of obtaining assessment data when that participant is not receptive. However, this approach remains underexplored, and so it is not immediately obvious how it might affect the collected data, either in quantity or quality. This paper aims to fill these research gaps.

To that end, we conducted a two-week field study with 15 main participants, who invited a total of 82 of their peers to provide data about them when they were not receptive, and then tested how well the peers' assessment data agreed with the main participants' own assessments. The three types of data we requested from both these participant groups, commonly sought in MA studies, were the main participants' 1) current locations (e.g., [8]), 2) activity types (e.g., [40]), and 3) emotional statuses (e.g., [12]). The field study *per se* was guided by two research questions:

RQ1 How many assessment responses could be obtained from the main participants' peers when they themselves were not receptive?

RQ2 How confident were the peers in their own assessments, and how well did their assessments agree with the main participants' own?

Additionally, with the aim of informing future users of our proposed MA approach about what kind of peers they should encourage their participants to recruit, we asked:

RQ3 What kind of peers, and how many of them, should be invited to participate in peer-assisted MA studies?

Our results showed that 96% of the occasions on which a main participant was non-receptive yielded at least one peer-assessment response; and in 42% of the MA questionnaires that received at least one peer response, at least one such response was given with high confidence. Moreover, 70% of the high-confidence responses agreed with the main participants' own responses. We also revealed that both the peers' confidence in their answers, and the agreement rate between such answers and those given by the main participants, varied according to both question type and the peers' characteristics. In addition, the nature of the peer-participant relationship and the frequency of such dyads' face-to-face meetings affected their agreement rate for location. The key contributions of this paper are as follows:

- It demonstrates that inviting peers to help answer MA questionnaires about the main participants' locations, activities, and emotions could increase the quantity of data obtained without diluting its reliability, provided that confidence information is collected in the peers' questionnaires.
- It shows that peers had differing levels of confidence in the accuracy of their answers to different types of questions, and that peer-participant agreement rates also varied by question type.
- It links specific peer characteristics to higher agreement rates for certain types of questions.
- It shows that the expected quantity of responses provided by different sizes of peer group also varied by question type.

These findings can serve as a useful reference for researchers determining peer-recruitment criteria for studies incorporating MA, as well as the approximate numbers of peers they ought to involve based on their research settings and purposes.

2 RELATED WORK

2.1 Mobile Assessment Studies

As MA becomes increasingly popular in various fields, methods for improving its data quality and/or quantity have attracted considerable research interest. As briefly noted above, longer study durations are ordinarily seen as allowing researchers to collect more data, but they can also lower participants' compliance [9] and willingness to respond [24, 36]. For this reason, Van Berkel *et al.* [42] suggested that a duration of two weeks was more suitable than longer periods. Likewise, although the typical aim of MA studies is the collection of data on a broad range of activities and situations that collectively depict the patterns of individuals' lives [22, 46], answering large numbers of daily questionnaires can impose undue burdens on participants [44]. Klasnja *et al.* [20] suggested that prompting participants five to eight times per day may be optimal.

Collecting data using different sampling strategies may cause different biases [22]. Researchers have used various scheduling mechanisms in MA, including signal-, interval-, and event-triggered ones [51], and Van Berkel *et al.* [45] showed that scheduling types affected participants' response rates and accuracy differentially. Lathia *et al.* [22] found that a random time-based method would create bias toward collecting data from people's most frequently visited contexts, while single sensor-based strategies were dependent on when the target event occurred. To alleviate contextual bias, a mixture of time-based and cognition-aware scheduling could be used [44]. A common combination is signal- and event-triggered scheduling, which enables researchers to capture experiences that pertain to specific times/events as well as those that occur throughout the day [42].

On the other hand, participants may be willing to provide responses at some moments but not others, and indeed, to be highly selective about this (e.g., [15, 28]); and Hormuth [15] argued that such selectivity would tend to cause bias. Our proposed strategy of inviting our main participants' peers to contribute to an MA study has the potential to mitigate this type of bias, by collecting information about a participant even when he/she is not receptive to questionnaires.

2.2 Increasing Data Quantity in Mobile Assessment

Response rate, also known as compliance rate [45], a general indicator of data quantity, is calculated by dividing the number of

completed questionnaires by the total prompts issued during a given study [42]. Various ways of boosting response rates have been developed, including visualization and gamification elements. Hsieh *et al.* [17] found that giving feedback to respondents after they completed questionnaires effectively increased their response rates. Van Berkel *et al.* [43] showed that a gamified condition outperformed a non-gamified one in terms of both data quality and quantity. Other approaches have included easing data entry, such as by replacing the regular smartphone-unlocking process with a microtask [39], and allowing participants to enter their answers directly into alert dialogues [47]. To address the issue of people rarely carrying their phones in certain environments, such as inside their own houses, Paruthi *et al.* developed a situated self-reporting system to collect participants' in situ stress, sleepiness, and activities in home environments [30]. They found that collecting responses via a smartphone plus one other designated device yielded more responses than would have been possible using just one or the other.

To avoid questionnaires being delivered at moments when users are unreceptive, researchers have attempted to use data from phones' sensors to predict opportune moments for notifications [31, 52] as well as questionnaires [13, 45], and other forms of data collection [48]. For example, Van Berkel *et al.* [45] found that scheduling questionnaires at phone-unlock moments yielded a high response rate. However, predicting opportune moments is difficult because of the many factors involved, only some of which might be discernible via their phones [25]. We expect that inviting MA study participants' peers to answer questionnaires about them will complement existing approaches to data-quantity maximization.

2.3 Peer-assisted Mobile Assessment

Berrocal *et al.* [3] argued that responses from the main participants' close friends or family members could contribute worthwhile data to EMA. However, they recruited such peers solely for the purpose of improving the accuracy of EMA responses about stress, based on an assumption that their main participants could not always reliably provide such information about themselves. We hope to extend that prior research by showing that peers' involvement can also increase data quantity and, depending on the peers' characteristics and numbers, data reliability also.

3 METHODOLOGY

In light of prior MA studies' findings that compliance gradually decreases over time [24, 36], we decided to divide our two-week study period into halves, and only ask the peer participants to provide assessments in the second week, so that we could observe whether compliance declined in that week as compared to the first, in line with theory [24, 36]. As a baseline for comparison, we included a second group of main participants, who completed the study on their own without the assistance of any peers in either week. However, despite this quasi-experimental design, it should be borne in mind that our primary focus was on answering the three research questions regarding the quantity and quality of peers' assessment data, and not on comparing main participants' compliance across groups. Below, we describe main-participant and peer recruitment, the MA questionnaires, and the study procedure.

3.1 Main Participant and Peer Recruitment

We recruited 27 main participants via posts in a Taiwanese subject-recruitment Facebook group followed by snowball sampling. Their ages ranged from 20 to 34, and all but two were undergraduate or graduate students. Because it was not possible to foresee or control how many (if any) or what kind of peers the main participants would be willing or able to recruit, we did not randomly assign the main participants into the with-peers and without-peers conditions. Instead, the 15 participants (seven males, eight females) who reported that they would each be able to recruit at least five peers to the study were instructed to do so; and the other 12 (six males, six females), who were less confident about their prospects of recruiting peers, were not instructed to do so, and thus formed the baseline without-peers group for comparison.

All participants and peers were required to have a smartphone with the LINE [1] messaging application, through which MA prompts would be delivered. A total of 82 peers (42 males, 40 females), with an average of around five (5.47) per main participants (Min: 5; Max: 9) were invited by the 15 main participants in the with-peers group. Of these peers, 38 were main participants' classmates; 29, their non-classmate friends; nine, their family members; and six, their significant others (SOs). The peers' ages ranged from 20 to 54, but the majority (n=70) were between 20 and 24. Compensation for participating in the study was US$36 and US$18 for the main participants and their peers, respectively.

3.2 Mobile Assessment Design

MA questionnaires were delivered via a customized chatbot using a LINE's API [2] service (hereafter referred to as the MA chatbot). The delivery of MA prompts was configured by a backend server built on the Heroku cloud platform [3]. Text-based chatbots have been identified in prior research as a promising means of gathering high-quality quantitative data, an advantage ascribed to their interactivity [19]. In this case, a further advantage of using a chatbot was that it did not require our main participants or their peers to install a stand-alone phone app, and thus was not limited to any specific phone-operating system [44]. All that was needed to receive MA prompts was to add the MA chatbot account as a contact in the messaging service. However, because no specialized research app had been installed on the main participants' or peers' phones, our inferences about good times at which to deliver MA prompts relied on the Google Calendar API [4]. Calendar information has previously been found effective as a means of predicting its users' interruptibility, without the need to consult sensor data [2, 16]. We therefore requested that all main participants arrange their daily events on Google Calendar during the study, so that MA prompts would be less likely to be delivered when they were busy. To further avoid disrupting their lives, we sent no MA prompts outside the hours of 10:00 AM to 10:30 PM, as recommended by previous studies [31, 42]. For the purpose of delivering MA prompts, each scheduled event on the main participants' calendar, as well as each period of time without any events scheduled (within the 12.5 hours of the day that

[1] https://linecorp.com/zh-hant/
[2] https://developers.line.biz/en/services/messaging-api/
[3] https://www.heroku.com
[4] https://developers.google.com/calendar/

the study ran), was identified as a discrete time block. Blocks lasting four hours or longer were divided into two-hour sub-blocks. All prompts were delivered at a randomly selected moment within 20 minutes after the end of each block, in line with prior research practices for labeling an immediately preceding activity/event (e.g., [7]). A minimum interval of one hour was also interposed between any two MA prompts. In practice, these criteria meant that the number of prompts per main participant per day ranged from six to eight; and every MA prompt sent to a main participant was also sent to all members of their peer group at the same moment.

3.3 Mobile Assessment Questionnaires and End-of-day Questionnaires

3.3.1 Mobile Assessment Questionnaires. Each MA questionnaire asked the main participant to report their location, activity, and emotion over the immediately preceding hour, in that order; and that person's peer-group members were asked for the same information about the main participant, based either on direct knowledge or on guesswork. To mitigate the response burden (e.g., [1]), we included predefined options for location and activity. As our recruitment pool composes mostly students, during a pilot study, we adjusted the options for location and activity by adding frequent responses they provided us with, such as library. Table 1 presents the answer options we used in this study. For emotion, in line with prior studies [21, 37], we included the eight dimensions shown in Table 1, each rated on a five-point Likert scale. Because we assumed that peers would often be unsure of their answers, we asked them to rate their confidence in each response they provided, an approach adopted from Berrocal and Wac [3]. Specifically, confidence was rated on a five-point Likert scale ranging from 4=very high to 0=very low. We did not include a "don't know" option, and made it clear in the study instructions that peers should report complete lack of knowledge/opinion as "0". The main participants and their peers were Mandarin speakers, so all questions were in Mandarin. Figure 1 (Left) presents mock-ups of the main-participant and peer versions of the MA questionnaire, translated into English.

3.3.2 End-of-day Questionnaire for Peers. We also wanted to know how often peers interacted with the participants who had recruited them, because we assumed that the frequency of such interaction might affect peers' confidence ratings. We also assumed that future researchers who involve peers in MA studies would prefer them to be able to provide reliable information over time, and direct interaction between peers and main participants would be a straightforward path to achieving this. Thus, the end-of-day questionnaire – delivered at 10:30 PM via the MA chatbot – sought a response for each time block to arrive at a detailed "snapshot" of the frequency of such interaction (Figure 1 (Right)). Only responses received before midnight that day were considered valid.

3.4 Study Procedure

All main participants in both the with-peers and without-peers group came to our lab to attend a training session, during which the researchers explained the study process, obtained access to the subjects' Google Calendar events, and instructed them to add the MA chatbot as a LINE friend. Additionally, the participants in the with-peers group completed a simple self-assessment about their peers. Specifically, they estimated how familiar each of their chosen peers had been with their locations, activities, and emotions in recent days, on a three-level scale comprising low, medium, and high familiarity.

Table 1: Answer options in each mobile assessment questionnaire, by dimension

Order	Dimension	Answer options
1	Location	Home, dormitory, office, classroom, library, store, restaurant, outside, gym, transportation, clinic, other (*choose one*)
	Confidence *	Five-point Likert scale (0-4)
2	Activity	In bed, doing leisure activities, shopping, eating, commuting, working, attending a meeting, studying, exercising, visiting doctor, other (*choose one*)
	Confidence *	Five-point Likert scale (0-4)
3	Emotion	Worried, angry, stressed, sad, bored, relaxed, happy, anxious (*each rated on a five-point Likert scale (0-4)*)
	Confidence *	Five-point Likert scale (0-4)

Note. *=only asked of peers.

Figure 1: (Left) Mock-ups, i.e., screenshots translated into English, of the MA Chatbot as it appeared to our users, including (a) the main-participant version and (b) the peer version. (Right) Mock-up of the end-of-day questionnaire.

Because the peers were distributed across a variety of cities, we contacted them via email rather than asking them to come to the lab. After adding the MA chatbot account as a LINE friend, they watched a three-minute video tutorial on how to respond to it, and afterwards were told to feel free to ask their main participant the answers to MA questions about him/her. This was because, at times when main participants are not receptive to MA prompts, they may still be selectively receptive to their peers' inquiries [23, 25]. Thus, we assumed that peers' proactive queries would, to a certain extent, increase our chances of obtaining main participants' self-assessments, as well as more accurate peer assessments. Also, inspired by Scollon *et al.* [28], who suggested that participants' personalities might affect compliance, the peers were also asked to complete the Chinese Big Five Personality Inventory (CBF-PI) [50].

Lastly, the main participants and peers were all told that, to help ensure the collected data's accuracy, it was best that they respond to any MA prompt within 30 minutes. However, we could not actually prevent them from responding later, due to a limitation of the chatbot, i.e., that messages remained in its chat window indefinitely, until removed by users or the operating system.

After the end of the second week of the field study, final instructions on how to uninstall the chatbot were sent out via the chatbot and email. At that point, we also asked all main participants and peers to provide informal feedback regarding their experience of peers helping to provide assessment data during the study. Compensation was then transferred to the subjects' bank accounts. The study was approved by our university's Human Ethics Committee, and was conducted in April 2019.

3.5 Measures

3.5.1 Response Rate. As recommended by Scollon *et al.* [28], MA responses that were delayed by 30 minutes or more were deemed invalid. A given MA prompt was deemed to have been responded to provided that at least one response was obtained, either from a main participant or any peer. Thus, the receipt of multiple responses by the same MA prompt did not increase the response count.

3.5.2 Response Agreement. Since we could not be sure that the main participants' responses regarding their own whereabouts, activities and emotions were correct or truthful, we did not use the term *accuracy*, but instead *agreement*, to indicate how often peers' answers were consistent with the main participants' own. We defined a peer as *reliable* if their responses achieved a high agreement rate. In the two multiple-choice dimensions, location and activity, the peer's answer had to be identical to the main participant's for us to deem agreement to have occurred. Thus, agreement in each of these two dimensions was a binary variable, with a value of either true or false. For emotion, on the other hand, since each response contained eight values, each from a different five-point Likert scale [21, 36], we treated a response as a vector and calculated the correlation between the response from the peer and that from the main participant. Raw agreement value was then calculated by Pearson correlation, and from that, we created a binary variable for emotional agreement, whereby a value of .7 or above was deemed true, and lower values false, following Ratner [33].

3.5.3 Meeting Frequency. We did not label each peer MA response with whether the relevant peer had met the main participant within the corresponding time period, because we could not know whether such physical meetings took place before or after the MA prompt was sent. Instead, we established that the top 25%, the median, and bottom 25% of meeting frequencies were 1.3, 0.2, and 0 times per day, respectively. This allowed us to generate four mutually exclusive categories of weekly meeting frequency: *two or more times per day*, *once per day*, *once every several days*, and *zero*.

4 RESULTS

We collected a total of 5,311 MA responses from all main participants in both conditions and the peers in the with-peers condition. Of these, 3,771 were responded to within 30 minutes. The main participants in the with-peers group received 6.9 prompts per day (SD=0.2), yielding an overall response rate 64.5%, with 63.7% in the first week and 65.2% in the second. Due to the structure of the study, as noted above, the average number of prompts per day received by peers was identical; however, their overall response rate was 55.0%. The participants in the without-peers group received 6.7 prompts per day (SD=0.2), and their overall response rate was 58.2%, with 61.1% in the first week, and 55.3% in the second. The peers' overall response rate to the end-of-day questionnaires was 80.7%. Slightly more than two-fifths of peers (41.5%; n=34) asked their associated main participants for status information, and the first quartile (top 25%) of them asked an average of up to 2.6 times per day.

4.1 Peers' Assessment Data Considerably Increased Data Quantity

4.1.1 Contribution of Peers' Assessment Data when Main Participants were Not Receptive. In answer to **RQ1**, on 96% of the occasions on which a participant was not receptive to a given MA prompt (i.e., did not respond to it within 30 minutes), at least one of his/her peers responded to that prompt. If we go one step further, and deem an MA prompt as "responded to" regardless of who responded, the overall response rate rises to 99%: i.e., more than one and a half times higher than the 65% response rate posted by the main participants alone. When we compared this overall response rate against that of the without-peers group in the same week, the difference was even more stark: 99% vs. 55%.

Next, we examined whether such response-rate differences held true if we only considered those peer assessments that were given with a high level of confidence. We found that, in 42% of the cases when a participant was not receptive, at least one of his/her peers provided a high-confidence assessment. The response rate represented by this combination of high-confidence peer responses and the main participants' own responses was 80%, or 1.3 times higher than that of the with-peers group's main participants alone in the first week, and 1.5 times the response rate of the without-peers group in the second week.

4.1.2 Main Participants' Response Rate. To assess how peer involvement affected the main participants' compliance, we ran a mixed-effects regression model with the independent variables *group* (i.e., with-peers vs. without-peers), *week* (i.e., week 1 vs. week 2), and the interaction of the two. We found an interaction effect between *group* and *week* on participants' response rate (see Table 2).

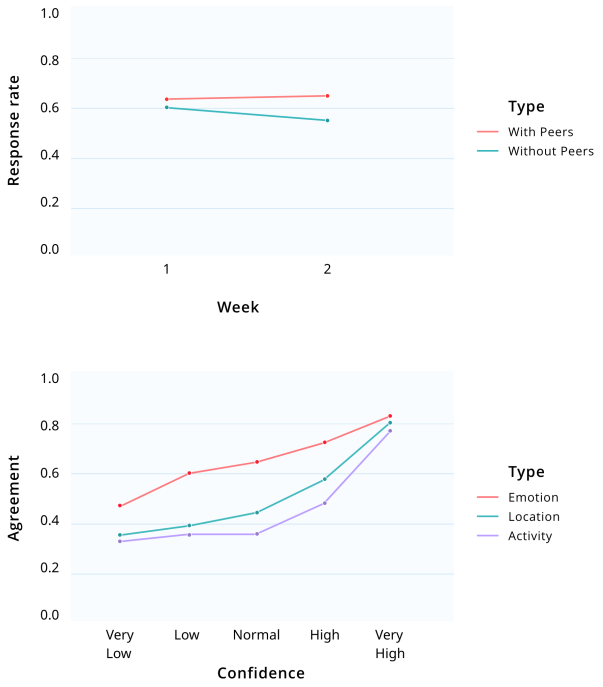

Figure 2: (Left) Week-on-week change in the average response rates of the main participants in the with-peers group and the without-peers group. (Right) Relationships of peers' confidence levels to peer/main-participant agreement levels, by item type.

Specifically, as Table 2 and Figure 2 (Left) show, the response rate of the without-peers group declined in week 2 (61%→55%), in line with prior research [24, 36]. Interestingly, however, the response rate of the main participants in the with-peers group did not decline (64%→65%).

We also learned from post-study feedback that, by asking for information about their statuses, peers sometimes reminded or motivated main participants to answer questionnaires that they might have ignored otherwise.

4.2 Confidence and Agreement among Question Types

To answer **RQ2**, we investigated the agreement and confidence rates associated with each of our three question types. The confidence rates, from highest to the lowest, were for questions about location (73%), activity (69%), and emotion (63%). Peers were very confident (i.e., rated "4") in their answers to these questions 35%, 29%, and 12% of the time, respectively, and confident (i.e., rated "3"), 26%, 28%, and 35% of the time. However, the agreement rates showed a quite different pattern: with emotion answers having the highest level of peer/main-participant agreement (69%), followed by location (62%), and activity (53%). Notably, the agreement rates for responses given with a confidence rating of "4" were similar across all three types of question (location: 81%; activity: 78%; emotion: 83%), but agreement rates for those responses with a confidence rating of "3" varied sharply (location: 58%; activity: 48%; emotion: 73%). In other words, peers were most likely to be confident when answering location questions, and least confident when answering emotion ones; and yet, their answers to emotion questions were the most likely to agree with the main participants'. Their answers regarding activity, meanwhile, were the least likely to agree with the participants'. We will discuss these discrepancies further in section 5.2, below.

Table 2: Non-standardized coefficients of mixed-effects logistic regression models predicting main participants' responsiveness (i.e., responded or did not respond) with a random effect to account for each participant.

Main Participants' Response		
Conditional R^2	0.2161	
	Est.	*p*
(Intercept)	0.72774	**0.00445** **
Week		
week 2	0.08334	0.49659
week 1	0 r	
Group		
without-peers	-0.21651	0.57015
with-peers	0 r	
Week * Group		
*week 2 * without-peers*	-0.35186	**0.04568** *
*p<.05 **p<.01 ***p<.001 r : reference		

Figure 3, which illustrates the distribution of peers' agreement rates, shows that these were lower for activity (red) and higher for emotion (green). This was mainly because the correlation values of most peers' emotion responses were greater than .7 (blue), despite their having the lowest confidence in such responses. Location (purple) also emerged as a relatively easy topic for peers to predict. The top location categories in terms of agreement were "office" and "dormitory"; and the top agreement categories for activity were "leisure activities" and "attending a meeting".

4.3 Preferred Peer Characteristics and Numbers

4.3.1 Relationship Type and Meeting Frequency. To answer **RQ3**, we explored the factors underlying peers' MA response data by building regression models aimed at achieving the best prediction of the agreement levels for each of our three question types. As Table 4 shows, we found main effects of relationship type and meeting frequency, but not perceived familiarity, suggesting its minor role in peers' response rate compared to the two former factors. These results indicate that how peers were related to main participants who invited them and how frequently they met with each other both impacted how often the peers' responses were consistent with the participants' own responses.

Specifically, among all the responses received from peers, responses received from SOs regarding participants' location (88.9%)

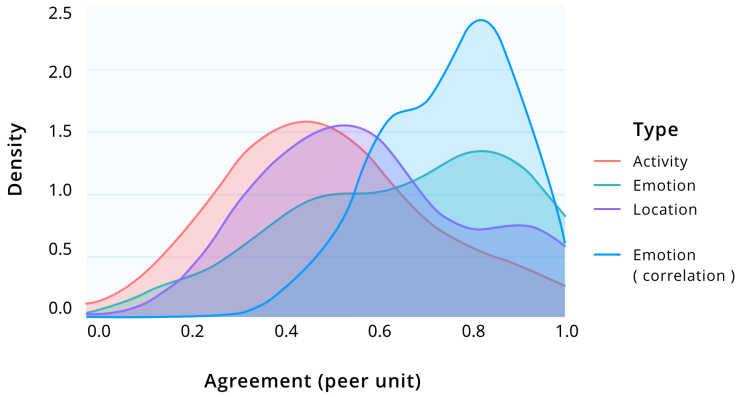

Figure 3: Density estimates of the agreement rates between peers and main participants, by question type.

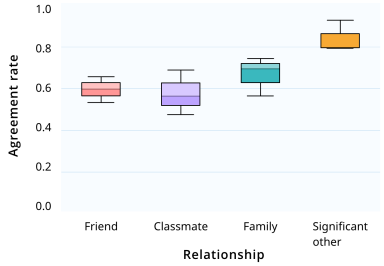

Figure 4: Distribution of agreement rates by peer/main-participant relationship type.

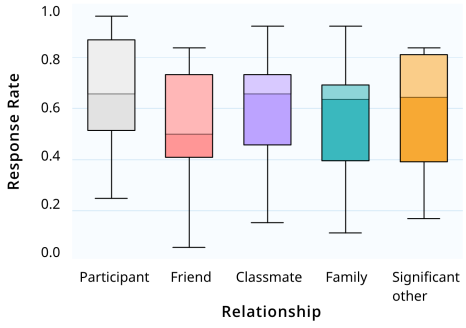

Figure 5: Distribution of response rates by peer/main-participant relationship type.

and activity (77.1%) were the most likely to agree with main participants' own responses among the relationship types, as shown in Table 3. The parallel location and activity figures for other peer types were only: 67.0% and 50.2% for family members; 55.2% and 45.8% for classmates; and 57.9% and 51.1% for friends. A similar pattern were also found from the angle of individual-level. That is, we calculated each individual peer's agreement rate, of which the distribution by relationship type is shown in Figure 4. It also shows that SOs' responses noticeably agreed more often with participants' own responses than the other relationship types. To confirm this effect, we ran logistic regression on peers' responses to examine if significant differences exist among relationship types in agreement likelihood. The differences between SOs and classmates and friends, respectively, were statistically significant (location: SO vs. classmates, Z=4.28, p<.001; SO vs. friends, Z=3.60, p<.001; activity: SO vs. classmates, Z=3.09, p=.002; SO vs. friends, Z=2.31, p=.02). In addition, the difference between family members and classmates was also significant (location: Z=3.41, p<.001; activity: Z=2.07, p=.04).

Unlike agreement rate, interestingly, we did not find noticeable differences in response rates across relationship types (family member: M=55.8%, SD=30.1%; SO: M=57.4%, SD=30.0%; classmate: M=58.3%, SD=22.4%; friend: M=50.3%, SD=22.9%), as shown in Table 3. From the individual-level angle, as shown in Figure 5, there seemed to be some differences among relationship types, with friends seemly were the least responsive category. However, due to the large variances within each category, as the figure shows, our logistic regression result – with the dependent variable being the binary outcome of whether an MA prompt was, or was not, responded to by a peer – does not show a main effect of relationship type on any type of question. This implies that every type of relationship included peers who were responsive and unresponsive to MA prompts, respectively, and such individual differences seemed to play a larger role than relationship type did. As a result, it seemed that relationship type influenced agreement more than it influenced response rates.

On the other hand, meeting-frequency categories were better predictors of high vs. low agreement for location and activity than for emotion. Specifically, those peers who met their main participant more often than twice per day were the most likely to provide location and activity responses that agreed with his/her own, i.e., 80% and 67%, respectively. Interestingly, this also implies that peers who met main participants frequently in person found the latter's activities more difficult to predict than their locations. There were no significant differences in the prediction of emotion-question

Table 3: Average response rate and agreement rates of peers for each question type, by peer-relationship type. N indicates the number of peers in that category.

	SO (N=6)	Friend (N=29)	Family (N=9)	Classmate (N=38)
Response rate	57.4% (SD=30.0%)	50.3% (SD=22.9%)	55.8% (SD=30.1%)	58.3% (SD=22.4%)
Location agreement rate	88.9% (SD=9.7%)	57.9% (SD=21.2%)	67.0% (SD=19.9%)	55.2% (SD=19.0%)
Activity agreement rate	77.1% (SD=14.8%)	51.1% (SD=23.1%)	50.2% (SD=21.4%)	45.8% (SD=17.2%)
Emotion agreement rate	78.7% (SD=18.9%)	66.6% (SD=25.2)%	68.1% (SD=21.1%)	65.3% (SD=22.2%)

agreement across our four meeting-frequency categories. Unexpectedly, the lowest agreement rate by meeting-frequency category was not posted by the "Zero" group, indicating that those peers who did not see the main participant at all during the second week of the study could still answer questions about his/her status with some accuracy.

Finally, as Table 4 also shows, there was a strong negative correlation between location agreement and peers' scores on the Big Five personality trait "openness to experience". That is, peers with high openness scores made numerous unconfident and wrong guesses about whether the main participant was at home or somewhere else: with nearly half of responses being low-confidence, and 80% failing to agree with the main participants' own responses. Messiah *et al.* [27] suggested that factors significantly associated with unanswered prompts included higher scores for "novelty seeking", which is closely related to "openness to experience" (see [11]). No similar effects were observed for activity or emotion, possibly because the agreement levels for such questions were generally low and high, respectively, which could have masked inter-peer differences. Lastly, the personality trait "conscientiousness" was positively correlated with activity-response agreement. This could have been because peers who were more persistent/responsible were more likely to answer relatively difficult questions. However, more research is needed to further investigate this relationship.

4.3.2 Relation of Response Quantity to Peer-group Size. When non-receptive to MA prompts, the main participants *ipso facto* did not provide MA responses, so it was not possible to calculate agreement rates for those occasions. Thus, the question of how many reliable responses we should expect to receive from peer groups of differing sizes when the main participant is non-receptive could not be precisely answered based on agreement levels. Therefore, we considered two alternative methods. The first was to estimate this quantity based on the number of responses expected when participants *did* provide a response. Specifically, for each main participant, we first calculated the average probability of obtaining at least one MA response from a given number of peers that agreed with the participant's own MA response. For example, a total of 10 possible pairs of peer responses could emerge from a group of five peers (C_2^5); thus, we calculated the probability of obtaining at least one MA response from two peers for each of these 10 outcomes, and averaged the probabilities of such outcomes. Then, we calculated the overall average of each participant's averaged probabilities across all participants, and obtained the results shown in Figure 6 (Left). The greatest change in the expected amounts was observed when the number of peers changed from one to two; and starting from three peers, we would expect an 80% chance of receiving an MA response to location and emotion questions from at least one peer. However, using this approach to estimate the expected quantity of peer responses when main participants are non-receptive might result in overestimation, given the aforementioned important role of peers' direct inquiries to the main participant in the former's answering process.

The second approach was to estimate the expected quantity via confidence. Given that the peers rated their confidence in their own responses even to questions that the main participants did not answer, we were able to estimate how many of these responses would likely have resulted in agreement (had there been anything to agree with), based on the known agreement rates associated with each confidence level. This approach, of course, assumes that the agreement rates are stable at each level of confidence, irrespective of whether the main participants are receptive. Figure 6 (Center) illustrates the probabilities of obtaining at least one response of at least a certain level of confidence. Specifically, when a given main participant's network comprised more than two peers, the probability of having at least one peer's answer rated at normal confidence or above (i.e., "3" or "4") was close to 1.0. However, the likelihood of at least one peer's answer to an emotion question being rated with very high confidence ("4") was considerably lower than in the other two question-type categories, as the same figure also shows.

Figure 6 (Right) shows the expected quantity of MA responses for each type of question. On average, networks comprising two, three, four and five peers led respectively to 0.38, 0.47, 0.53 and 0.58 high-confidence responses per MA prompt. Expected quantity was lowest when one considered only responses given with very high confidence. Even when five peers responded, the expected quantity of very-high-confidence responses was below 0.4 for all types of questions, reflecting that the average probability of any peer giving such a rating to any response was just 25%, or 12% in the case of emotion questions. Although the agreement rates at normal ("2") and high ("3") confidence levels were both lower, it should be borne in mind that participants were much more likely to assign such ratings to their responses than "4" ratings. This meant that the expected quantities at those confidence levels were quite high. Not unexpectedly, emotion questions had the lowest expected quantity at the very-high-confidence level, given that peers were least often confident about their answers to emotion questions. Importantly, however, such questions had the highest expected quantity at normal confidence. It is also noteworthy that the expected quantities of "3"+ answers to both location and activity questions were nearly equivalent to those of "2"+ answers in the same two question-type categories. This was because the differences in the expected quantities of "3"+ answers we observed grew as the number of peers increased (1⟶2: +57%; 2⟶3: +22%; 3⟶4: +13%; 4⟶5: +10%).

The two kinds of estimation described above led to quite different expected-quantity trends. Figure 6 (Left) showed that involving two peers would achieve an expected quantity of 0.7 responses per MA prompt across all types of question, while involving three peers would achieve 0.8. However, Figure 6 (Right) indicates considerably lower expected quantities of responses with "3"+ confidence: i.e., not quite 0.4 per MA prompt from a group of two peers, and 0.5 from three peers. While we cannot know which of these two methods best represents the ground truth, caution dictates that we regard the

Table 4: Non-standardized coefficients of mixed-effects logistic regression models predicting agreement between the main participants' and their peers' answers, by question type, with a random effect to account for each participant.

	Location		**Activity**		**Emotion**	
Conditional R^2	0.2326		0.1455		0.2263	
	Est.	*p*	*Est.*	*p*	*Est.*	*p*
(Intercept)	0.1512	0.6024	-0.2673	0.3045	1.0960	**0.00155** **
Conscientious	0.0198	0.7692	0.1441	**0.03447** *	-0.0549	0.4721
Openness	-0.1715	**0.038655** *	-0.1492	0.0776	0.0327	0.7219
Agreeableness	0.0311	0.6612	-0.0066	0.9221	0.0462	0.5687
Neuroticism	-0.1004	0.1479	0.0264	0.6909	-0.0536	0.5017
Extraversion	-0.0470	0.5982	-0.0722	0.3769	0.0121	0.9048
Relationship						
Significant Other	1.7975	**<0.001** ***	0.9980	**0.00201** **	-0.1563	0.6824
Friend	0.2669	0.1594	0.1767	0.3143	0.0799	0.7358
Family	0.8079	**<0.001** ***	0.4747	**0.03828** *	-0.2155	0.2467
Classmate	0 r		0 r		0 r	
Familiarity						
Low	-0.2116	0.4364	-0.1510	0.6102	-0.0834	0.7377
Mid	-0.3419	0.0998	-0.1439	0.5109	-0.1991	0.4242
High	0 r		0 r		0 r	
Meeting Frequency §						
High	0.9665	**<0.001** ***	0.5670	**0.03784** *	-0.0293	0.8922
Normal	0.3964	0.1082	0.4457	**0.04783** *	0.2653	0.3783
Low	0.1493	0.3709	0.2368	0.1481	-0.4026	**0.04102** *
Zero	0 r		0 r		0 r	
*p<.05 **p<.01 ***p<.001 r : reference						

§ *High* refers to a frequency of two or more times per day; *Normal* to once per day; *Low* to once every several days; and *Zero* to no meetings

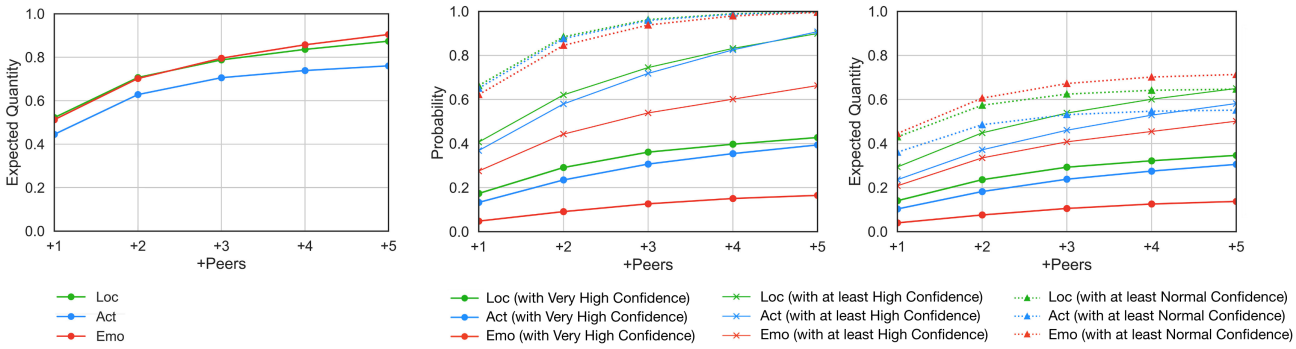

Figure 6: Probability of obtaining agreement between peers' and main participants' answers. (Center) Probability of obtaining at least one peer answer with at least normal confidence (i.e., 2+ on a scale of 0-4). (Right) Likelihood of obtaining reliable peer answers to complement main participants' non-responses.
***Note.* Loc=location; Act=activity; Emo=emotion**

first method as more likely to produce overestimates of response quantities.

5 DISCUSSION

5.1 Enhancing Data Collection during Participants' Unreceptive Moments

Prior researchers have proposed numerous approaches for increasing MA response rates while maintaining data quality, but none has been without its drawbacks. Our results have shown 1) that allowing MA participants to invite their peers to contribute information about them is a feasible approach to enhancing data collection, especially – but not exclusively – at moments when the former are not receptive to MA prompts; and 2) that data quality can be assured by asking such peers about their levels of confidence in the accuracy of their own responses. Even when only responses rated with a high or very high confidence level were considered, the proposed peer-assisted approach yielded an average of more than three additional

responses for every eight responses from the participants themselves. Taken together, this implies that in the real-world scenario of data collection that leverages MA, such as collecting labeled activity data (e.g.[6]), practitioners and researchers may consider recruiting participants who are already socially connected at a group level and employ this peer-assisted approach to ensure data quantity. Interestingly, however, the higher overall number of responses we received from the with-peers group than from the without-peers group was *not* solely due to the peers' contributions, but also to an elevated response rate from the main participants during the second week of the experiment, i.e., when their peers joined them in it. This finding was diametrically opposed to our assumption that the main participants' compliance with MA prompts would decrease in the second week. We made this assumption due both to compliance's widely acknowledged tendency to decrease over time [24, 36], and to our intuition that our main participants would expect their peers to "cover for" them. From feedback, we learned that this unexpected increased-compliance phenomenon occurred because their peers asked them questions about MA questionnaires, and in some cases, directly reminded them to answer them. Such discussions indeed became a motivator for them to sustain their response rates, which increased marginally over the course of the study, from 64% to 65%, in contrast to the without-peer group's response rate, which fell, in line with expectations [24, 36].

As we expected, the more confident peers were in their responses, the more likely such responses were to agree with those given by the main participants. This implies that confidence ratings could help researchers tell which responses are likely to be reliable vs. unreliable. However, confidence alone is not always adequate: notably, peers' responses at the very high confidence level ("4") were inconsistent with the main participants' responses slightly more than 20% of the time. Berrocal *et al.* [3] argued that individuals might not be capable of reliably answering questions about their own current emotional states. Nevertheless, two of our three question types did not involve emotion, and still only had an 80% agreement rate when peers were very confident in their answers. In any case, it must be remembered that it was infeasible within the constraints of our study to know whose answers were factually correct when main participant/peer disagreement occurred.

5.2 Question-type Effects on the Reliability of Peers' Answers

Peers' confidence levels and agreement rates both varied considerably according to the type of question. Peers were most confident in their location responses, but those responses were not actually the most likely to agree with the main participants' ones. Conversely, the responses peers were least confident in, i.e., regarding emotion, were not in fact the least likely to agree with the participants' own. These discrepancies can be attributed to two causes.

The first is that humans' emotions, unlike physical locations and activity types, tend to be distributed on a spectrum, and to be subject to frequent – if usually small – hour-by-hour fluctuations [5, 49]. As such, even though each main participant's peers had little if any knowledge of their real-time emotional status, it was still relatively easy for peers to provide emotion-related answers that were highly correlated to the main participant's own answers, provided only that none of the parties thought the main participant's current emotional state was extreme. However, peers' relatively low confidence in their emotion-related answers could have reflected an understanding that emotions are transitory and easily affected by moment-to-moment experience, and thus less likely to be known as more time elapses since the parties' most recent face-to-face interaction. This probably also explains why we did not observe a main effect of either relationship type or meeting frequency on the agreement levels of answers to emotion-related questions. In contrast to emotion, location was relatively constant within a given block of time, and thus more easily predicted by peers who knew the main participants' schedules. This was also reflected in the main effects of relationship type and meeting frequency on location-question agreement levels. Although activity and location data are often closely interrelated, and perhaps especially so in university settings where buildings' functions tend to be highly specialized, questions about activity were more difficult for peers to answer correctly than location ones were. However, the lower agreement rates for these two question types than for emotion-related questions was rather unexpected. It could have occurred because the main participants were engaged in multiple activities and/or at several locations, but only reported the major ones in their own responses. Also, people can be reluctant to make detailed online disclosures of contextual information that might enable others to infer their daily routines or habits [8]. In light of these two factors, peers' lack of confidence in their activity and location responses is perhaps less surprising.

The second likely cause of variation in peers' confidence levels and agreement rates across question types is related to how we asked the questions. For instance, if we had asked peers to rate the likelihood that the main participant would stay at a certain place or engage in particular activities, or to choose an emotional label for them rather than using a Likert scale, we would undoubtedly have seen different response distributions, and thus, probably, different agreement rates. But, by the same token, we expect that researchers using similar approaches to ours when asking location-, activity-, and emotion-related questions will obtain results broadly similar to ours.

5.3 Characteristics of High-agreement Peers

In our regression analyses, we included fixed effects of three categorical variables – relationship type, perceived familiarity, and meeting frequency – that we assumed researchers could obtain for their MA studies via questionnaires. Our results suggest that relationship type was a good predictor of the agreement level of peers' answers regarding location and activity, but not emotion. Overall, SOs and family members were the two categories of peers most likely to provide responses consistent with the main participants' own, with SOs answers being the more consistent of the two. However, no particular class of peer was notably more likely than any other to provide the same answers to emotional questions as the main participants did. Meeting frequency, meanwhile, was a better predictor of location-related agreement than of the other two agreement types.

Peers who met with the main participant two or more times per day were more likely to achieve location and activity agreement

than those who did not meet with the main participant at all during the study period. This result is not surprising, insofar as peers who met a main participant frequently had more chances to directly observe them. Interestingly, however, those peers who *never* met their main participants were not the *least* likely to achieve agreement with their answers. This could have been because the main participants disclosed their status via other communication channels, such as social media, messaging apps or phone calls [14]. That is, a category of "online/remote peers" might occasionally be able to provide accurate information about main participants. It therefore might be worthwhile for future MA studies to investigate whether the frequency with which main participants use computer-mediated communication such as texts and social-media posts are correlated with agreement levels.

5.4 Effects of Peer-group Size

As Figure 6 shows, the enhancement of response quantities associated with larger peer-network sizes was most marked when participants nominated only one vs. two peers, and gradually diminished as the numbers of peers grew. While a similar pattern has emerged during usability testing [29], we cannot draw any firm conclusions about how many peers would be "enough" for an MA study. Among the numerous other factors that might be involved, and which thus merit further investigation, Figure 6 highlights three: the researchers' desired minimum response rate, their desired minimum agreement rate, and their specific questions. Researchers will also confront an important trade-off when deciding the minimum confidence threshold for a peer's response to be deemed acceptable. That is, accepting only responses given with very high confidence could help ensure data quality; yet, peers might assign their answers such high confidence ratings only rarely, meaning that the majority of peer data would be eliminated. Conversely, lowering the threshold of confidence would tend to increase the quantity of data, but entail more risk of data unreliability. Taking all of this into account, our general recommendation is that researchers utilize high- *and* very high-confidence responses, to ensure that data quantity is sufficient, as this is, after all, our proposed approach's central aim. Importantly, this acceptance of slightly lower data quality in exchange for higher data quantity will be more worthwhile when the number of peers per main participant is small (i.e., between one and three). This is because, with larger peer-group sizes, the probability of at least one peer per question rating their answer with high or very high confidence is also greater, and this effect will naturally lead to data being collected on a wider range of MA questionnaires. In short, while we cannot propose a "right" number of peers to involve in MA, Figure 6 may be helpful to researchers seeking to determine such a number for their own MA studies. They might also usefully adopt our approach of estimating peers' responses' agreement rates at different levels of confidence, and then using those rates to determine confidence thresholds.

5.5 Study Limitations

This preliminary investigation of peer-assisted MA is subject to numerous limitations, and leaves many questions unanswered. First, the majority of our participants were postsecondary students. This could have rendered their locations and activities relatively more foreseeable by their peers than would otherwise have been the case. Therefore, whether our findings are generalizable to other populations with less predictable routines is unclear.

Second, while not all the participants we recruited were willing to invite any peers to join this study, some who were willing to do so could not convince any peers to accept their invitations. This meant that our with-peers group was subject to a self-selection bias toward those with larger numbers of willing peers, which will have limited the diversity of participants and peers alike. Future research should try to expand the pool of participants to more diverse populations.

Third, to calculate agreement rates, we treated the main participants' answers as a gold standard. However, Berrocal *et al.* [3] have argued that peers' responses might be more accurate than the participants' own, at least in the case of stress-related questions. Nor could we have done anything to prevent our main participants from intentionally providing socially desirable but untrue answers, which would have caused their peers' truthful answers disagree with theirs. Moreover, we did not analyze agreement *among* peers' responses, which in future might provide another useful means of assessing the truthfulness of all responses. Such an inter-peer agreement metric could be especially useful in cases where the main participants have not provided any responses.

Fourth, we only considered three types of questions, whereas in reality, MA is used in a much wider variety of research domains. More importantly, we had specific ways of asking peers about location, activity, emotion, and confidence; and, as mentioned earlier, this resulted in specific patterns of response distribution, which in turn may have influenced agreement rates and self-reported confidence. Although it is commonplace in MA studies to ask about participants' locations and activities via multiple-choice questions, and about their emotions via scales, researchers may sometimes need to choose different approaches, including but not limited to open-ended questions. Therefore, whether the results of our study would be replicable if other answer formats were used would need to be established through further research. If we had included an option such as "Don't know", our observed agreement rates and data quantity might have been different too, since peers are likely to choose a label only when they have a certain level of confidence. Also, our questionnaires asked main participants and peers to label a previous hour, which may mean that our results are not generalizable to ESMs that ask for real-time, in-situ experiences.

Fifth, we did not collect any information on interactions between a peer and a main participant at the moment the former responded to an MA prompt, such as whether they were co-located, or whether the peer had asked the main participant about how to answer. The reason for this was that asking about these matters in each MA might have rendered the overall study experience too burdensome for peers. Nevertheless, such information might be useful to know.

Finally, this paper methods were almost exclusively quantitative. Thus, future scholars of peer-assisted MA should consider collecting the main participants' and peers' perceptions and attitudes.

6 CONCLUSION AND DIRECTIONS FOR FUTURE WORK

Responses in a traditional MA study can only be obtained when its participants are receptive to MA prompts. In this paper, we have shown that inviting peers to provide information about the main participants' statuses is an alternative means of capturing data at moments that are inopportune for the latter group of individuals. We found that inviting peers improved overall data quantity, even when only high-confidence answers were considered. Unexpectedly, peers' participation also seems to have motivated the main participants to sustain their compliance, i.e., to continue responding regularly in the second week, in contrast to the control-group members, whose compliance slackened over time. However, peers expressed very different levels of confidence in their answers to different kinds of questions, and this was a major factor in the observed variation in the rates of agreement between main-participant and peer responses. Relationship types and meeting frequencies were good predictors of such agreement, but only for certain types of questions. The ideal size of a peer group, therefore, would depend heavily on what questions the researchers plan to ask, and how much of a tradeoff they are willing to make between data quality and data quantity. In sum, we have shown that peer-assisted MA is a promising approach; but numerous questions about it remain unanswered, and deserve further and deeper investigation.

ACKNOWLEDGMENTS

We thank our main participants and peer participants. This work was generously supported by the Ministry of Science and Technology, R.O.C. (MOST 110-2636-E-007-004, MOST 109-2218-E-009 -016, MOST 107-2218-E-009 -030-MY3).

REFERENCES

[1] Aino Ahtinen, Minna Isomursu, Ykä Huhtala, Jussi Kaasinen, Jukka Salminen, and Jonna Häkkilä. 2008. Tracking Outdoor Sports – User Experience Perspective. In *Ambient Intelligence*, Emile Aarts, James L. Crowley, Boris de Ruyter, Heinz Gerhäuser, Alexander Pflaum, Janina Schmidt, and Reiner Wichert (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 192–209.

[2] Anja Bachmann, Robert Zetzsche, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. Identification of Relevant Sensor Sources for Context-Aware ESM Apps in Ambulatory Assessment. 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* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 265–268. https://doi.org/10.1145/2800835.2800944

[3] Allan Berrocal and Katarzyna Wac. 2018. Peer-Vasive Computing: Leveraging Peers to Enhance the Accuracy of Self-Reports in Mobile Human Studies. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 600–605. https://doi.org/10.1145/3267305.3267542

[4] Daniel Buschek, Sarah Völkel, Clemens Stachl, Lukas Mecke, Sarah Prange, and Ken Pfeuffer. 2018. Experience Sampling as Information Transmission: Perspective and Implications. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 606–611. https://doi.org/10.1145/3267305.3267543

[5] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 3 (March 2018), 20 pages. https://doi.org/10.1145/3191735

[6] Yung-Ju Chang, Gaurav Paruthi, Hsin-Ying Wu, Hsin-Yu Lin, and Mark W. Newman. 2017. An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-word settings. *International Journal of Human-Computer Studies* 102 (2017), 81 – 102. https://doi.org/10.1016/j.ijhcs.2016.11.001 Special Issue on Mobile and Situated Crowdsourcing.

[7] Ian Cleland, Manhyung Han, Chris Nugent, Hosung Lee, Shuai Zhang, Sally Mcclean, and Sungyoung Lee. 2013. Mobile Based Prompted Labeling of Large Scale Activity Data. *Ambient Assisted Living and Active Aging Lecture Notes in Computer Science* (2013), 9–17. https://doi.org/10.1007/978-3-319-03092-0_2

[8] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. 2005. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 81–90.

[9] Delphine Courvoisier, Michael Eid, and Tanja Lischetzke. 2012. Compliance to a Cell Phone-Based Ecological Momentary Assessment Study: The Effect of Time and Personality Characteristics. *Psychological assessment* 24 (01 2012), 713–20. https://doi.org/10.1037/a0026733

[10] Mihaly Csikszentmihalyi and Reed Larson. 2014. *Validity and Reliability of the Experience-Sampling Method*. Springer Netherlands, Dordrecht, 35–54. https://doi.org/10.1007/978-94-017-9088-8_3

[11] Filip De Fruyt, L. Wiele, and Kees van Heeringen. 2000. Cloninger's Psychobiological Model of Temperament and Character and the Five-Factor Model of Personality. *Personality and Individual Differences* 29 (09 2000), 441–452. https://doi.org/10.1016/S0191-8869(99)00204-4

[12] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying Emotional States Using Keystroke Dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 715–724. https://doi.org/10.1145/1978942.1979046

[13] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Designing An Experience Sampling Method for Smartphone based Emotion Detection.

[14] Jeffrey A. Hall and Nancy K. Baym. 2012. Calling and texting (too much): Mobile maintenance expectations, (over)dependence, entrapment, and friendship satisfaction. *New Media & Society* 14, 2 (2012), 316–331. https://doi.org/10.1177/1461444811415047 arXiv:https://doi.org/10.1177/1461444811415047

[15] Stefan E Hormuth. 1986. The sampling of experiences in situ. *Journal of Personality* 54, 1 (1986), 262–293. https://doi.org/10.1111/j.1467-6494.1986.tb00395.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1986.tb00395.x

[16] Eric Horvitz and Johnson Apacible. 2003. Learning and Reasoning about Interruption. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 20–27. https://doi.org/10.1145/958432.958440

[17] 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* (Seoul, Korea) *(UbiComp '08)*. Association for Computing Machinery, New York, NY, USA, 164–167. https://doi.org/10.1145/1409635.1409657

[18] Ashish Kapoor and Eric Horvitz. 2008. Experience Sampling for Building Predictive User Models: A Comparative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 657–666. https://doi.org/10.1145/1357054.1357159

[19] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 86, 12 pages. https://doi.org/10.1145/3290605.3300316

[20] Predrag Klasnja, Beverly L. Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E. Hudson. 2008. Using Wearable Sensors and Real Time Inference to Understand Human Recall of Routine Activities. In *Proceedings of the 10th International Conference on Ubiquitous Computing* (Seoul, Korea) *(UbiComp '08)*. Association for Computing Machinery, New York, NY, USA, 154–163. https://doi.org/10.1145/1409635.1409656

[21] Peter Kuppens, Nicholas B. Allen, and Lisa B. Sheeber. 2010. Emotional Inertia and Psychological Maladjustment. *Psychological Science* 21, 7 (2010), 984–991. https://doi.org/10.1177/0956797610372634 arXiv:https://doi.org/10.1177/0956797610372634 PMID: 20501521.

[22] Neal Lathia, Kiran K. Rachuri, Cecilia Mascolo, and Peter J. Rentfrow. 2013. Contextual Dissonance: Design Bias in Sensor-Based Experience Sampling Methods. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 183–192. https://doi.org/10.1145/2493432.2493452

[23] Hao-Ping Lee, Kuan-Yin Chen, Chih-Heng Lin, Chia-Yu Chen, Yu-Lin Chung, Yung-Ju Chang, and Chien-Ru Sun. 2019. Does Who Matter? Studying the Impact of Relationship Characteristics on Receptivity to Mobile IM Messages. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*

(Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 526, 12 pages. https://doi.org/10.1145/3290605.3300756

[24] Derrick C. McLean, Jeanne Nakamura, and Mihaly Csikszentmihalyi. 2017. Explaining System Missing: Missing Data and Experience Sampling Method. *Social Psychological and Personality Science* 8, 4 (2017), 434–441. https://doi.org/10.1177/1948550617708015 arXiv:https://doi.org/10.1177/1948550617708015

[25] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2858036.2858566

[26] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. 2015. Ask, but Don't Interrupt: The Case for Interruptibility-Aware Mobile Experience Sampling. 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* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 723–732. https://doi.org/10.1145/2800835.2804397

[27] Antoine Messiah, Olivier Grondin, and Gaëlle Encrenaz. 2011. Factors associated with missing data in an experience sampling investigation of substance use determinants. *Drug and Alcohol Dependence* 114, 2 (2011), 153 – 158. https://doi.org/10.1016/j.drugalcdep.2010.09.016

[28] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. *Experience Sampling: Promises and Pitfalls, Strength and Weaknesses*. Springer Netherlands, Dordrecht, 157–180. https://doi.org/10.1007/978-90-481-2354-4_8

[29] Jakob Nielsen and Thomas K. Landauer. 1993. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 206–213. https://doi.org/10.1145/169059.169166

[30] Gaurav Paruthi, Shriti Raj, Seungjoo Baek, Chuyao Wang, Chuan-che Huang, Yung-Ju Chang, and Mark W. Newman. 2018. Heed: Exploring the Design of Situated Self-Reporting Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 132 (Sept. 2018), 21 pages. https://doi.org/10.1145/3264942

[31] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 897–908. https://doi.org/10.1145/2632048.2632062

[32] Tauhidur Rahman, Mi Zhang, Stephen Voida, and Tanzeem Choudhury. 2014. Towards Accurate Non-Intrusive Recollection of Stress Levels Using Mobile Sensing and Contextual Recall. ICST. https://doi.org/10.4108/icst.pervasivehealth.2014.254957

[33] Bruce Ratner. 2009. The correlation coefficient: Its values range between 1/-1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing* 17, 2 (2009), 139–142. https://doi.org/10.1057/jt.2009.5

[34] Aki Rintala, Martien Wampers, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2018. Response Compliance and Predictors Thereof in Studies Using the Experience Sampling Method. *Psychological Assessment* (11 2018). https://doi.org/10.1037/pas0000662

[35] Saul Shiffman, Arthur Stone, and Michael Hufford. 2008. Ecolocial Momentary Assessment. *Annual review of clinical psychology* 4 (02 2008), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[36] Paul J. Silvia, Thomas R. Kwapil, Kari M. Eddington, and Leslie H. Brown. 2013. Missed Beeps and Missing Data: Dispositional and Situational Predictors of Nonresponse in Experience Sampling Research. *Social Science Computer Review* 31, 4 (2013), 471–481. https://doi.org/10.1177/0894439313479902 arXiv:https://doi.org/10.1177/0894439313479902

[37] Gerasimos Spanakis, Gerhard Weiss, Bastiaan Boh, Lotte Lemmens, and Anne Roefs. 2017. Machine learning techniques in eating behavior e-coaching. *Personal and Ubiquitous Computing* 21, 4 (Aug 2017), 645–659. https://doi.org/10.1007/s00779-017-1022-4

[38] Arthur Stone, Saul Shiffman, and Audie Atienza. 2007. *The Science of Real-Time Data Capture: Self-Reports in Health Research*.

[39] Khai N. Truong, Thariq Shihipar, and Daniel J. Wigdor. 2014. Slide to X: Unlocking the Potential of Smartphone Unlocking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3635–3644. https://doi.org/10.1145/2556288.2557044

[40] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. 2018. ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 554, 12 pages. https://doi.org/10.1145/3173574.3174128

[41] Niels van Berkel, Matthias Budde, Senuri Wijenayake, and Jorge Goncalves. 2018. Improving Accuracy in Mobile Human Contributions: An Overview. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 594–599. https://doi.org/10.1145/3267305.3267541

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

[43] Niels van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 107 (Sept. 2017), 21 pages. https://doi.org/10.1145/3130972

[44] Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. 2019. Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 51, 12 pages. https://doi.org/10.1145/3290605.3300281

[45] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. 2019. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies* 125 (2019), 118 – 128. https://doi.org/10.1016/j.ijhcs.2018.12.002

[46] Simone J W Verhagen, Laila Hasmi, Marjan Drukker, J van Os, and Philippe A E G Delespaul. 2016. Use of the experience sampling method in the context of clinical trials. *Evidence-Based Mental Health* 19, 3 (2016), 86–89. https://doi.org/10.1136/ebmental-2016-102418 arXiv:https://ebmh.bmj.com/content/19/3/86.full.pdf

[47] Aku Visuri, Zhanna Sarsenbayeva, Niels van Berkel, Jorge Goncalves, Reza Rawassizadeh, Vassilis Kostakos, and Denzil Ferreira. 2017. Quantifying Sources and Types of Smartwatch Usage Sessions. https://doi.org/10.1145/3025453.3025817

[48] Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 14 pages. https://doi.org/10.1145/3098279.3098532

[49] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[50] YAO Shu-qiao WANG Meng-cheng, DAI Xiao-yang. 2011. Development of the Chinese Big Five Personality Inventory(CBF-PI) III: Psychometric Properties of CBF-PI Brief Version. *Chinese Journal of Clinical Psychology* 4 (2011).

[51] Ladd Wheeler and Harry T. Reis. 1991. Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality* 59, 3 (1991), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1991.tb00252.x

[52] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5346–5360. https://doi.org/10.1145/3025453.3025946