Published in final edited form as:

Proc ACM Interact Mob Wearable Ubiquitous Technol. 2024 November ; 8(4): . doi:10.1145/3699735.

Ask Less, Learn More: Adapting Ecological Momentary Assessment Survey Length by Modeling Question-Answer Information Gain

JIXIN LI,
Northeastern University, USA

ADITYA PONNADA,
Northeastern University, USA

WEI-LIN WANG,
University of Southern California, USA

GENEVIEVE F. DUNTON,
University of Southern California, USA

STEPHEN S. INTILLE
Northeastern University, USA

Abstract

Ecological momentary assessment (EMA) is an approach to collect self-reported data repeatedly on mobile devices in natural settings. EMAs allow for temporally dense, ecologically valid data collection, but frequent interruptions with lengthy surveys on mobile devices can burden users, impacting compliance and data quality. We propose a method that reduces the length of each EMA question set measuring interrelated constructs, with only modest information loss. By estimating the potential information gain of each EMA question using question-answer prediction models, this method can prioritize the presentation of the most informative question in a question-by-question sequence and skip uninformative questions. We evaluated the proposed method by simulating question omission using four real-world datasets from three different EMA studies. When compared against the random question omission approach that skips 50% of the questions, our method reduces imputation errors by 15%−52%. In surveys with five answer options for each question, our method can reduce the mean survey length by 34%−56% with a real-time prediction accuracy of 72%−95% for the skipped questions. The proposed method may either allow more constructs to be surveyed without adding user burden or reduce response burden for more sustainable longitudinal EMA data collection.

Author Contact Information: Jixin Li, li.jix@northeastern.edu, Northeastern University, USA.

Additional Keywords and Phrases:

Ecological momentary assessment; experience sampling; survey length; question informativeness; user burden; machine learning

1 INTRODUCTION

Ecological momentary assessment (EMA), also known as the experience sampling method (ESM), is a method for repeatedly measuring self-reported behaviors and states using mobile devices in real-world settings [10, 61]. Users receive prompts on their mobile devices to complete self-reported surveys consisting of a set of questions; the questions may ask about in-situ behavior and/or states, especially those that passive sensors cannot reliably measure yet (e.g., perceived fatigue [11] and pain [34]). EMAs may reduce recall biases compared to retrospective recall surveys and enhance ecological validity; measurements can be repeated many times in a day, thus capturing temporal changes within individuals over time [56]. Because of these benefits, EMA is frequently used to gather in-situ self-reports by health behavior researchers [8, 25, 57] and by human-computer interaction researchers [5, 16, 24] to study free-living behavior, and to gather person- and context-specific data to inform just-in-time adaptive interventions (JITAI) in digital health applications [3, 29, 53].

In the past two decades, researchers have been increasingly interested in deploying EMA in longitudinal data collection, with temporally dense self-report [34, 68, 78], but sustaining long-term user engagement is difficult, partly because of heavy user burden [65]. A typical EMA protocol used in health or psychology could have eight prompts a day, with question sets up to 36 questions long and take 1–2 min to answer (e.g., [27]). The frequent requests for responses to lengthy surveys impose disruption in daily life, which cumulatively may lead to poor EMA compliance and inattentive responding [13, 14, 59]. Thus, improving data collection efficiency and minimizing user burden has been a critical challenge in EMA methodology research [19, 47].

One way to reduce burden might be to prompt surveys at less disruptive times, and previous studies have explored methods to predict such moments using passive sensing data; examples are prompting surveys during unproductive usage sessions [69], activity transitions [1], and breakpoints in social contexts [43]. This approach may, however, exacerbate selective non-response bias and limit the generalizability of the data to all contexts and situations [66]. Alternatively, a protocol might use random-time sampling but reduce the survey length to only those questions that provide valuable information at that moment. Researchers have recently proposed to deliver adaptive-length psychological assessment by integrating EMA with psychometric techniques from item-response theory (IRT) [21, 53]. These methods use IRT-calibrated sets of questions (i.e., ‘item banks’) to drive computer adaptive testing/measurement (CAT) whereby a subset of questions is selected from the item bank to assess a single construct efficiently. This method, while promising, is dependent upon the development of IRT-calibrated item banks [53].

Inspired by the prior work, we propose a question-selection method that can be used with random-time sampling and that does not rely on a pre-calibrated item bank. We define the potential information gain of obtaining the answer to a question as the prediction uncertainty of the question-answer, inferred from prior EMA answers in the study and participants' answers to other questions in the current EMA survey. The proposed method starts with building person-specific prediction models based on responses to the full-set EMA surveys in the first several weeks of longitudinal data collection. Then when a survey is initiated, questions from the question set are selected one-by-one, where at each step the question yielding the highest information gain is chosen. By strategically skipping the unselected questions that are likely to provide little new information given the questions already answered, user burden may be reduced, but with minimal information loss. The results from question omission simulations using real-world longitudinal EMA datasets suggest that our method could significantly reduce survey length while causing less information loss compared to random-question omission.

Our work makes three contributions:

- We present a practical method to shorten survey length with minimal information loss by quantifying the information gain of each question in a survey set using only prior responses to EMA survey questions. Unlike IRT-based CAT techniques that measure a single construct with fewer questions, our method can be applied to EMA surveys with one or multiple questions per construct, as long as the questions or constructs to be modeled are interrelated.
- We assessed the performance of the proposed method in both longitudinal observational and interventional EMA data collection settings. The simulation studies on four real-world EMA datasets show that our method effectively predicts responses for real-time monitoring, allowing us to skip 34–56% of questions while maintaining answer prediction accuracy of 72–95% for skipped questions with five answer options. Additionally, the method results in lower imputation errors for post-study analysis compared to a planned missing data design, which randomly omits questions to shorten surveys [58].
- We demonstrate the generalizability of the proposed method by testing it on four datasets from real-world longitudinal EMA studies, each with different study objectives, study populations, study durations, prompt schemes, and question sets. Specifically, we show the efficacy of employing the proposed method in EMA studies using time-based (e.g., report every hour or every day) and event-contingent (e.g., report after each interpersonal interaction) prompt schemes, asking questions about both mental state and behavior, and using single-item rating scales (rapid assessment for easily understandable constructs) and composite rating scales (precise assessment for complex constructs).

In the remainder of this paper, we review prior work on reducing user burden in EMA in Section 2. We outline our research questions and hypotheses in Section 3. In Section 4, we describe the main components of our proposed method. In Section 5, we describe the four real-world EMA datasets used for simulation studies, demonstrate the prompt question-selection process through examples, and we report the evaluation results of our method through a series of simulation studies using real-world longitudinal EMA datasets.

Finally, we discuss the implications of our results in Section 6 and possible directions for future research in Section 7.

2 RELATED WORKS

In this section, we review the problem of user response burden with EMAs and prior research aimed at reducing the burden using adjustments to prompt timing and survey length.

2.1 Impact of User Burden on EMA Studies

Long-term monitoring is critical for longitudinal studies that study behavior changes [12, 64] and mental disorders [48, 74], as well as developing digital health interventions [31]. The intensive longitudinal data collection made possible using EMA may provide deep insight into individuals' daily life experiences, but maintaining high-quality self-report data collection over time is challenging. Van Berkel et al. [65] conducted a survey of 110 experience-sampling papers published between 2005 and 2016. Most of these EMA studies (70.9%) ran for less than a month, with an average length of 32 days. Earlier researchers running diary studies [60] found that incomplete or incorrect data increased after two to four weeks. As a result, Van Berkel recommended a data collection duration of one to three weeks for EMA studies that study a phenomenon that requires multiple self-report surveys each day. With a one-month study duration on average, the average response rate of studies included in the survey [65] was 69.6% (i.e., the number of fully completed questionnaires divided by the number of questionnaires presented). This average response rate falls below the 75% threshold, above which samples are typically considered with high compliance and less prone to selection bias [62]. In sum, the user burden associated with EMA hinders longitudinal data collection using the method.

2.2 Computational Methods to Reduce EMA Response Burden

Three broad strategies can be used to reduce EMA response burden: improving the presentation of questions (e.g., enhancing text readability [9], using multiple modalities [6]), improving the timing of prompts (i.e., identifying less disruptive moments to prompt [35]), and reducing the need for self-report data (i.e., reducing survey length, number of surveys prompted per day, and observation duration) [65]. The timing-related and data-related design components of EMA are illustrated in Figure 1. In this section, we summarize previous work on using computational methods to improve the timing of prompts and reduce the length of surveys.

2.2.1 .Reducing Response Burden by Adjusting Prompt Timing.—Fogarty et al. [17] demonstrated that environmental sensors can be used to estimate human interruptibility as well as people can. Using manually simulated video and auditory sensors and machine learning models, they achieved an accuracy of 82.4% in detecting interruptibility, which was better than humans on the same task (76.9%). Subsequent studies have attempted to predict less disruptive moments for prompting using passive sensing data to enhance users' self-report response rates. For example, Obuchi et al. [41] found higher response rates (58% vs 50%) when prompts were delivered during activity-based breakpoints (e.g., from walking to stationary) detected by the Activity Recognition API in iOS. Similarly, Aminikhanghahi et al. [1] used unsupervised change point detection algorithms to identify activity transitions in real-time using smart home sensors (passive infrared motion and door sensors) and found a higher response rate during the transitions (84.26%) compared to random-time prompting (78.62%). Visuri et al., [69] prompted surveys during unproductive smartphone usage sessions after the phone screen was unlocked. Zhang et al. [79] investigated 'unlock journaling' (i.e., answering a question while unlocking the phone), which led to a higher frequency of reporting as well as decreased perceived intrusiveness. Park et al. [43] deferred notifications until breakpoints in a social context and thereby reduced the number of disruptive notifications by 54.1%.

Although finding opportune prompting times may help to decrease perceived burden and increase response rates, prompting only at moments with high response rates may exacerbate participant compliance bias and selective nonresponse bias [66]. Van Berkel et al. [66] analyzed four smartphone EMA studies and found considerable differences in survey compliance rates among participants. They found that participant compliance bias was significantly impacted by contextual factors detected using smartphones, such as time of day, weekday/weekend, screen status, last phone usage. A more recent study examined various contextual factors such as time, device use, physical activity and mobility and found most of them may be associated with non-response to prompts on smartphones and smartwatches [45]. In addition, nonresponses have been found to be associated with higher levels of stress and negative affect, resulting in biased parameter estimates [39]. These results suggest that participant response rates are not consistent across time and context, and prompting only at moments when participants are more willing to answer may worsen such biases in the data collected, potentially affecting the reliability of models built from those data [66].

2.2.2 Reducing Response Burden by Adjusting Survey Length.—

Although both sampling frequency and length of surveys contribute to participants' perceived burden [56], previous studies suggest that longer surveys may impact burden more negatively than high-sampling frequency [14, 28]. For instance, Morren et al. [37] reviewed 62 papers published from 1991 to 2006 on assessing daily pain using electronic diaries. Across these studies, the survey length varied from one item to 63 items per survey, and the number of diaries ranged from 1 to 10 per day. The regression analysis showed that diary length had a statistically significant negative association with compliance, while the number of diaries per day did not. Specifically, with each additional item in the survey, the compliance rate was reduced by 0.48%. In a more recent study, Eisele et al. [14] conducted randomized controlled trials with 150 participants to investigate how survey length and sampling frequency influence perceived burden, compliance, and careless responding. They found that longer surveys had statistically significantly lower compliance, higher momentary burden, and more careless responses than short versions of the surveys, and that increased sampling frequency only influenced retrospective burden through interaction with longer surveys.

Planned missing data designs, also known as random-question omission, have been used to reduce the length of surveys by randomly omitting survey questions [23, 58]. Silvia et al. [58] investigated the effect of applying two variations (matrix sampling and anchor test) of planned missing data designs to EMA research. For instance, the matrix sampling method randomly assigns a subset of questions for each prompt occasion. Simulation studies showed that the planned missing data designs yielded unbiased parameter estimates at the cost of higher standard errors against the complete-case sample.

Rather than dropping questions randomly, the information contribution of each question can be considered [22]. Schneider et al. [53] introduced just-in-time adaptive EMA (JITA-EMA) to classify a person's momentary state with high accuracy using a small subset of EMA questions. They extended the use of computerized adaptive testing (CAT) methods that are rooted in item response theory (IRT) to classify momentary states that inform timing decisions of JITAI treatment. The method uses an 'item bank' consisting of rigorously tested collections of questions that can be used to estimate the true value of a unidimensional construct of interest. Strategic question selection can be achieved by selecting the next question that is the most informative given the current estimate of the true value of the hidden construct; this question selection continues until a classification decision on users' momentary states can be made with sufficient confidence (e.g., 95% confidence interval of the estimated construct level did not include the cutoff for classification). In simulation studies, JITA-EMA with only two to three questions on average (out of the total 13 questions) per prompt achieved better classification accuracy than using a fixed five questions per prompt. The method leverages information value in survey question selection when calibrated IRT question sets are available and provides reliable estimates of tailoring variables for JITAI with reduced-length surveys.

3. REASERCH QUESTION

We extend the prior work by considering the following research question: In longitudinal data collection, without altering sampling schemes or requiring a calibrated item bank of questions, how might we quantify the information gain of each survey question and use this to deliberately shorten the length of EMA surveys, possibly reducing user burden while minimizing information loss?

We propose a method to explicitly predict the potential information gain of answers to survey questions by leveraging individuals' response patterns from prior answer history. An idiographic model can be fitted to an individual's response data during the longitudinal data collection that can encode predictive relationships between constructs of interest. The underlying assumption is that constructs being surveyed are, to some degree, interrelated (i.e., exhibit probabilistic dependency [20]), which is typically observed in EMA studies and digital health applications that assess interconnected aspects of human behaviors (e.g., social activities) and mental states (e.g., affect and feelings). Thus, a prediction model can be used to determine the uncertainty in predicting the current states of unobserved constructs, quantifying the potential information gain if the participant answers the survey question about a construct. Using information gain has been explored in active learning [54], where the goal is to find a small set of informative samples that can optimize the performance of prediction models. In this work, we apply these concepts to longitudinal EMA to achieve adaptive question selection at each survey prompt occasion.

We evaluate the proposed adaptive prompt-question selection method by simulating data collection using existing real-world smartphone EMA study datasets from prior work. Our hypotheses are:

- **H1**: As more questions of a survey are answered, the uncertainty about the remaining question answers will decrease based on past observations. Compared to random-question omission (i.e., planned missing data design), the proposed method, prioritizing the presentation of the most uncertain question, can accelerate the uncertainty reduction for the remaining questions and skip low-uncertainty questions with minimal information loss, resulting in higher real-time prediction accuracy for skipped questions.
- **H2**: Compared to random-question omission, the proposed method can skip the same total number of questions while incurring less information loss in self-report data collection, resulting in lower imputation errors for missing data.
- **H3**: The proposed method can be applied to different types of constructs (e.g., mental states and behaviors), prompt schemes (e.g., time-based or event-contingent), and question sets (e.g., single-item or composite rating scales) because each prompted survey is treated independently and the method has the flexibility in modeling interrelated constructs that measure different aspects of human behavior at different levels.

4 METHDOLOGY: ADAPTIVE EMA PROMTING

This section introduces the components of the proposed adaptive EMA prompting. We first describe how to estimate potential information gain when questions are answered based on prior answer history (Section 4.1). Then, we describe how to select questions based on the estimated information gain and when to stop asking questions to achieve adaptive survey lengths (Section 4.2).

4.1 Estimation of Potential Information Gain from Obtaiing Answer to an EMA Survey Question

To assess the potential information gain from obtaining answers to an EMA survey question, we leverage prediction models to learn the relationships between constructs from the past data and determine the uncertainty level of states of unobserved constructs based on observed constructs for each occasion.

4.1.1 Prediction Model on Question-Answer.—The adaptive EMA method requires building prediction models for users’ answers to survey questions. Although many models can achieve this, we use Bayesian networks (BN) in this work because they can be used to reason about uncertainty [15]. BNs can represent a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) and a set of node probability tables (NPTs). In the DAG, the nodes represent random variables (discrete or continuous) and the directed arcs represent causal or influential relationships. A directed arc connecting A→B means A is the parent node of B and the dependency can be quantitatively represented as the conditional probability P(B|A) in the NPT of node A. By making conditional independence assumptions between unlinked variables, a BN simplifies the calculation of the joint probability distribution. The inference process, based on Bayes' theorem [4], refers to information of observed variables (evidence) being propagated through the DAG to update prior probability distributions for every unobserved variable, in both forward (from parent nodes to child nodes) and backward directions (from child nodes to parent nodes). Because of these features, the BN models are widely used to reason with uncertainty in disease diagnosis and complex healthcare decisions involving multiple factors that influence each other [33].

We use a BN to support question-selection decisions. The BN can be used to reason about the uncertainty level of states of multiple variables that represent answers to survey questions. Because a BN can update uncertainties with incomplete data, it only requires a single model to estimate the probabilities of states for all unobserved question variables, based on the responses to any other set of question variables. In other words, when a participant answers a survey question, the prediction model can revise the posterior probability distributions for all the other survey questions that have not been answered. For example, when asked about stress levels, if a participant reports being “Quite a bit” or “Extremely” stressed, that report may make the model more certain about the state of constructs such as “happy” and “relaxed” if such relationships have been present in prior data. As more responses are entered, the model's predictions on the states of unobserved variables become more certain. The process of question selection may then stop at some predetermined uncertainty threshold using a stopping rule (discussed later in Section 4.2).

4.1.2 Intializtion and Continuous Updating of the Trediction Model.—

Before starting the question selection process, the method requires collecting participants' responses to full surveys for a short period to initiate the model training. This full-survey training data can be sourced from previous studies that measured similar constructs for the same user groups. Alternatively, the training data may be obtained from pilot studies or the early stages of longitudinal data collection, during which participants typically exhibit higher motivation to engage in EMA compared to later stages. If surveys include more constructs or have weak correlations between constructs, or if respondents are less consistent in their responses, more training data may be required. After initialization, the model can be used to skip questions and collect partial survey data. As data collection progresses, the model may be continuously updated with participant responses from all prior completed surveys. An alternative is to keep training the model using a shifting window of the most recent response data. In this work, we chose the former updating strategy because it ensures that the model benefits from the full breadth of participant responses up to a given point in time. When training the model, we used observed responses for presented questions and predicted responses for skipped questions from previously completed surveys. This approach, known as pseudo-labeling, is a semi-supervised learning technique that incorporates unlabeled data into supervised learning to improve prediction performance [42].

4.1.3. Estimation of Expected Information Gain—

The method allows survey constructs to be measured on either discrete or continuous scales. In this work, we used discrete options because our test data have discrete question answers. The state of a discrete random variable represents the participant's selected answer option to a question. For prediction models, the input is the states of observed variables, and the output is the estimated likelihood of all possible states of the unobserved variables. There are many ways to summarize the prediction uncertainty with a single quantity (e.g., the posterior probability of the most probable label, the margin between the first and the second most probable labels) [54]. In this work, we use entropy from information theory [55], which is often used as an uncertainty measure in machine learning. The function of expected information gain for a survey question is:

$$Expected\ \ information\ \ gain = H(X) = -\sum_i p(x_i)\log(p(x_i))$$

where X represent states of an unobserved variable, x_i ranges over all possible discrete states of the variable, $p(x_i)$ is the likelihood of the state predicted by the idiographic model, $H(X)$ is the entropy of the predicted likelihood of all possible states of the variable. Intuitively, a high value of entropy resulting from an even probability distribution means we are less certain about the outcome, and vice versa. For example, a binary variable with an answer probability (0.5, 0.5) has an entropy of 0, while the same variable with an answer probability (1.0, 0.0) has an entropy of 1 with a logarithm base 2. For representation simplicity, one can use a base set to the number of response options (e.g., in this work, we use a base 5 for 5-point rating scales and a base 3 for 3-point rating scales) to scale the range of entropy value to 0–1. The prediction uncertainty inferred from the model can then be used to quantitatively represent the potential information gain of a survey question to be presented.

4.2 EMA Survey Question Selection and Stopping Rules for Each Prompt Occasion

The question selection process of a hypothesized EMA survey is conceptually illustrated in Figure 2. The EMA survey question selection rule is to include the survey question with the highest information gain (estimated by prediction uncertainty) among all unknown questions. For each survey to be prompted, the first question is selected based on the prior distribution of the BN model because no survey question has been answered. Starting from the second question, all prior responses will be included as evidence to update the probability distribution for all other unknown questions. The updated probability distribution will then inform the selection of the next survey question. Each prompted question set is considered independently of prior question sets. Question selection continues until either all questions in a question set are answered, or a stopping rule ends question presentation.

We implemented three stopping rules:

- **Fixed length**: Questions are picked one-by-one based on the expected information gain until a fixed number of questions is reached for each prompt.
- **Variable length**: Questions are picked one-by-one based on the expected information gain until the prediction uncertainty of all unobserved questions is lower than a threshold.

- **Variable length with cap**: Similar to the variable-length strategy, except there is an additional constraint that no survey can be longer than a maximum number of questions.

5 SIMULATION STUDIES

In this section, we first describe the datasets used for simulation studies to evaluate the proposed method and the technical details of model building. Next, we discuss the insights we learn from the prompt decision process through a real-world survey example. We then present the results from a series of simulation studies that systematically evaluate the proposed method using four real-world EMA datasets with various data collection settings.

5.1 Datasets

We conducted question omission simulations on four datasets from three real-world longitudinal EMA studies. The comparison of EMA datasets can be found in Table 1. Dataset 1 and Dataset 2 were from the same study [46, 71] and were mainly used to test H1 and H2 on the feasibility and efficacy of our method; Dataset 3 and Dataset 4, from two EMA studies [18, 76] that differ from the first study in terms of study population, study duration, prompt scheme, and question set, were used to test H3 on the generalizability of our method.

5.1.1 Dataset 1 and 2: Daily and Hourly Surveys of a Year-Long EMA Study.—

We used the data collected as part of the Temporal Influences on Movements and Exercise (TIME) study [46, 71]. The goal of that study was to explore temporal factors that influence health behavior change and maintenance in young adults (ages 18–29 years) using EMA and mobile sensing. Participants answered self-report surveys using both EMA on smartphones and μEMA [28] on smartwatches for up to 12 months; in this work, we only used EMA data collected on the smartphones. A mixed sampling design was used in this study, from which we extracted two datasets: **Dataset 1 (daily-EMA) and Dataset 2 (hourly-EMA).**

The **daily-EMA dataset** consists of responses to end-of-day surveys prompted on all days before the participant's anticipated sleep time. The end-of-day question sets have 12 questions about different mental state constructs, taken directly or adapted from established measures [75]: affective and feeling states (*happy, energetic, relaxed, sad, fatigued, tense, stressed, frustrated, nervous*), attention (*focused*), self-control (*resist*), and productivity (procrastinate). For the daily-EMA, questions asked, "Over the last day, how [construct] did you feel?" to capture daily summaries, with response options on a five-point scale labeled "Not at all," "A little," "Moderately," "Quite a bit," and for the fifth item, either "Extremely" or "Very much so." Some other questions with different answer styles relating to sleep time and health behavior are less relevant to this work and were excluded from the analysis.

The **hourly-EMA dataset** consists of responses to surveys prompted every waking hour on days of 'burst periods.' Each month, there were two burst periods when participants answered the EMA surveys once an hour during the waking hours. Waking hours were set by the participants. Each burst period consisted of four days with two weekend days guaranteed. There was at least a one-week gap between the two burst periods each month. A participant who self-reported 16 h of wake and 8 h of sleep was expected to receive 15 prompts per burst-day. The hourly surveys on burst days used the same 12 questions as in the daily-EMA surveys with minor modifications on question text (i.e., “Right now, how [construct] do you feel?” instead of “Over the last day[…]”) to capture momentary reports.

The full list of survey questions can be found in Table 7 of the Appendix. Using all responses in the datasets, we examined the correlation matrices for measured constructs in both daily and hourly-EMA surveys (Appendix, Figure 9). Moderate to strong correlations were found between surveyed constructs, which indicates potential prediction relationships between each other.

Of 135 participants who completed the one-year study, we excluded 15 participants who responded to less than 250 daily-EMA surveys (i.e., response rate below ~70%) from the daily-EMA dataset to ensure data quality and sufficient data points for one-year analysis [62]. We also excluded one participant from both datasets who was reported in the dataset codebook as having admitted to inattentive responses in the exit-interview. Descriptive statistics about the daily and hourly-EMA survey datasets can be found in Table 2.

5.1.2 Dataset 3: Fixed-Time Surveys of a Two-Week EMA Study.—

This dataset was derived from an EMA study in which investigators aimed to explore the changes in mental health and social contact of college students during the outbreak of the COVID-19 pandemic in the Netherlands [18]. Participants answered EMA surveys at four fixed times (noon, 3 p.m., 6 p.m., and 9 p.m.) each day for 14 days. During the two-week study period, the Dutch government announced a series of strict rules on social distancing, group gatherings, and home quarantine. The study found that participants’ mental health and social behavior were significantly changed over the two weeks due to the release of these policies. Each survey contains 17 questions, including nine questions on mental health states and eight questions on social contact behavior and COVID-19-related behavior. We excluded the hunger question from the original survey because it was found to be independent of all other constructs in the learned DAG structure (discussed more in Section 5.2). The full list of questions can be found in Table 8 of the Appendix. The mental health questions, adapted from standardized scales, inquired about participants’ feelings over the past three hours, with five-point answer options ranging from ‘Not at all’ to ‘Extremely.’ The behavior questions, created by the researchers, asked participants to report the time they spent on different activities over the past three hours, using one of the five categories: “0 min,” “1–15 min,” “15–60 min,” “1–2 hours,” and “> 2 hours.” For both types of questions, responses were scored on a scale of 1 to 5, with 1 representing “Not at all” or “0 min” and 5 representing “Extremely” or “> 2 hours.” Of 79 participants in the original dataset, we excluded five participants who had missing data for five days or more from the dataset to ensure sufficient data for analysis.

5.1.3 Dataset 4: Event-Contingent Surveys of a Three-Week EMA Study.—

The study that generated this dataset explored how personality disorder affects the dynamics of romantic relationships [76]. Participants were recruited using a stratified design, ranging from minimal or no symptoms to a positive screen for personality disorders. The EMA study used an event-contingent prompt scheme where participants were asked to actively report their mood and interpersonal behavior on smartphones immediately after every interpersonal interaction for 21 days. The survey included 31 distinctive items of momentary affect adjectives from the extended version of the Positive and Negative Affect Schedule (PANAS-X [73]). Unlike the single-item scales used in Datasets 1–3, where one item is used to define each construct, this question set is a composite scale that consists of six multi-item subscales in a hierarchical structure: two general dimension subscales about the valence of the mood (positive affect and negative affect), and four specific affect subscales about the distinctive qualities of the negative affect (anxiety/fear, anger/hostility, shame/guilt, and depression/sadness). Although the positive affect subscale consists of a distinctive item set, the negative affect subscale shares 0–4 items with each specific negative affect subscale. The detailed item composition of each subscale can be found in Table 9 of the Appendix. Each question is rated on a scale of zero to four, with zero representing “Not at all” and four representing “Extremely.” The composite score of each high-level affect construct is created by the average score of multiple low-level affect items contributing to the high-level affect construct.

5.2 Building the Model

We used the *bnlearn* (v0.8.2) Python package [63] to learn the DAG structure and NPTs to build the BN model. First, we learned the DAG structure of constructs measured in surveys. Using score-based approaches, we chose the commonly used Bayesian information criterion (BIC) [40] as the score function to evaluate how well the model fits the data, and we used the hill climb search algorithm [50] to traverse the search space. Because the structure search algorithm for BN models requires a large amount of data, we used initial full-survey training data from all participants to learn the DAG structure. Separate DAGs were learned for each EMA dataset. Example DAG structure plots can be found in Figure 10 of the Appendix. Second, the initial full-survey training data from individual participants were used to learn the NPT parameters to obtain person-specific models. We estimated the conditional probability distributions of the individual variables using maximum likelihood estimation (i.e., relative frequencies).

The amount of initial training data needed is still an open question and, in this work, we tested how the amount of initial training data impacts method performance over time (see Section 5.4.4). For the initial full-survey training periods, the models used for Dataset 1 and 2 surveys were initially trained on data from one to three months at the beginning of the study. One month’s data includes about 30 days of daily EMA data in Dataset 1, but it only corresponds to about eight days of hourly EMA data in Dataset 2. To test whether our method can work with a small amount of training data, the model used for the Dataset 3 and 4 surveys was initially trained only on the data from the first one-to-two weeks of the study.

In this work, we examine the applicability of our method to different types of EMA surveys. We demonstrate two model-building scenarios using different EMA datasets. For surveys in Datasets 1–3, each single question item corresponds to a construct measured. For example, the question “Right now, how happy do you feel?” in the hourly survey measures the momentary happiness construct. In this scenario, we fitted the model directly on question score data and made prompt decisions on individual questions. For surveys in Dataset 4, multiple question items contributed to each high-level construct. For example, the scores of six question items were used to compute the composite score of the anxiety/fear construct (Table 9). We assume researchers are interested in using the composite scores of high-level affect constructs rather than the scores for individual question items. Therefore, in this scenario, we constructed the BN model to learn the relationships between composite scores of different high-level affect constructs. Each node in the DAG represents one of these constructs. As a result, when one affect construct was selected to be omitted, we could skip the entire question set contributing to that affect construct.

5.3 Example of the Prompt-Question Selection Process

Before discussing method performance, we present an intuitive example of prompt-question selection process using predicted information gain. Consider one example survey from an individual in Dataset 1. Using the first three months of daily EMA data for initial model training, we applied the proposed method to the surveys presented in the fourth month. We illustrate the step-by-step question selection process for a survey in Figure 3. One survey consisting of 12 questions on daily affect and feelings is presented to the participant. In a question-by-question sequence, the algorithm selects the question item with the highest prediction uncertainty level and obtains the answer to that question from the participant. The answer is used to update the question-answer prediction (integer 0–4) and uncertainty level (float 0–1, in parenthesis) of all unselected survey question items. For example, the first column ("0 (None)") of the figure shows the initial answer prediction and uncertainty level of all survey items (inferred from prior answer history) when no questions have been answered. The question on 'Resist' is then selected. The cell values of all other unselected questions are shown in the second column ("1" question answered) and reflect the new levels of uncertainty after incorporating information from the first answer. Green cells indicate when the predicted responses match the ground-truth responses; red cells indicate where the prediction would fail. The bottom two rows of the figure show the average prediction uncertainty and accuracy of all unselected questions each time a new question is answered.

From this example, we can observe several important phenomena. First, as more questions were answered, the average uncertainty level of unknown questions decreased from 74% to 9%, and the overall average prediction accuracy increased from 42% to 100%. This suggests that the question selection strategy based on information gain could reduce the overall uncertainty level so that the model becomes more certain about unknown questions as more evidence/information is observed. Second, we observed that predictions for three questions (tense, stressed, and nervous) were corrected after observing responses to the previous highly uncertain questions. At step seven, the level of being tense increased from 0="Not at all" to 1="A little" after observing the increase of the participant's procrastination level from originally predicted 0="Not at all" to 1="A little." Similarly, the model predicted the participant to be more stressed (from 0="Not at all" to 1="A little") after the participant reported more frustration (from 0="Not at all" to 1="A little") at step six; the model predicted the participant to be more nervous (from 0="Not at all" to 1="A little") after the participant reported to be not relaxed (from 1="A little" to 0="Not at all") at step nine. This result suggests the model may achieve accurate predictions on unselected questions after enough evidence is observed. In fact, after nine questions were answered by the participant (step nine in Figure 3), the answers to the three remaining questions were all correctly predicted by the model. Thus, for this prompt moment, there exists an optimal prompt strategy that stops the survey after nine questions are observed in the order suggested by the model, reducing the original survey length by three questions yet losing no information. Thus, this example demonstrates the feasibility of reducing the survey length while minimizing information loss by using information gain to select survey questions.

We also observe that the average accuracy rate is not monotonously increasing; it has small fluctuations for the first six steps. In this example, the first bits of information show the participant was in a state that is not commonly seen (feeling in control and focused quite a bit but not happy at all). Therefore, the model was highly uncertain at first (average uncertainty level entering a plateau), but when at step 7 and 8 the participant reported procrastinating and being stressed, the average uncertainty level decreased. This inspired us to experiment with a variable-length stopping rule based on uncertainty level rather than a fixed number of questions (see Study 2 in Section 5.4.2).

5.4 Results of Simulation Studies

In this section, we present the results from a series of studies that simulated question omission using our method on real-world EMA datasets. To test our hypotheses, we first investigated the impact of the proposed question-selection method (Study 1). We then evaluated the model performance in both interventional EMA data collection (Study 3, real-time prediction) and observational EMA data collection (Study 6, imputation errors). To test the generalizability of the proposed method, we replicated Study 3 and Study 6 on two additional datasets with different study settings (Study 7). Additionally, to inform real-world implementation, we examined the impact of using different types of stopping rules (Study 2), the impact of initializing the prediction model using different lengths of initial full-survey training periods (Study 4), and the impact of measurement precision (Study 5).

5.4.1 Study 1: Impact of Question Selection Using Information Gain.—

Study 1 aims to examine the feasibility of using information gain in question-selection decisions using Dataset 1 (**H1**). Motivated by the example in Section 5.3, we analyzed how answering each question selected by our method impacts two factors: uncertainty and accuracy of predicting question answers. We compared the changing trend of these two factors between our method and random-question selection. We also examined the impact of the proposed method on question presentation order.

The relationship between the number of answered questions, prediction uncertainty, and prediction accuracy on all daily surveys from the fourth month of data collection across all participants is shown in Figure 4 (left). First, by selecting each question based on information gain, the average uncertainty level of unknown questions decreases, and the average prediction accuracy increases as more questions are answered by participants; the prediction uncertainty has a strong negative correlation with prediction accuracy ($r = -0.48$,

$p < 0.05$), indicating the feasibility of using uncertainty to estimate information gain. Second and more importantly, compared to the results of selecting each question at random, our method reduced the overall prediction uncertainty level and increased the overall prediction accuracy of the unknown questions to a greater degree for each step (i.e., steeper slopes). These results indicate the effectiveness of our method in reducing prediction uncertainties on unknown questions by selecting more informative questions, which enables more efficient prompting by asking fewer questions.

A side-effect of question selection based on information gain is the permutation of question presentation order. We found only 36.8% of surveys (SD = 0.24) are repetitive in presentation order on average, although some questions were more likely to be selected in the earlier steps than others (Appendix, Figure 11). The varying presentation order of questions may reduce the chances of participants providing neutral or inattentive responses in repeated surveys [70].

5.4.2 Study 2: Comparison of Stopping Rules.—In Study 2, we compared three different stopping rules (i.e., fixed-length, variable length, variable length with cap) using Dataset 1. We varied the values of parameters associated with different stopping rules to examine how they influence the number of questions skipped and prediction accuracy on unselected questions. For the variable-length rule, we selected a range of stopping thresholds from 0.001 to 1.0. For the fixed-length rule, we tested the number of prompted questions from none to all (0 to 12). The variable-length with caps rule used caps of 6, 8, and 10. The same testing set was used for evaluation as indicated in Section 5.3.

Figure 4 (right) shows the comparison results. By varying the stopping thresholds of the variable-length rule, one could control the trade-off between the question skipping percentage and prediction accuracy. With a lower (i.e., stricter) stopping threshold for skipping questions, the model would skip a smaller percentage of questions on average (from 100% to 3.8%) with a higher prediction accuracy (from 55.7% to 87.6%); and the optimal performance (skipped percentage = 12.5%, prediction accuracy = 87.6%) was achieved with a threshold of 0.2. Similarly, with the fixed-length strategy, prediction accuracy increases as the percentage of skipped questions decreases. As expected, the variable-length rule in general outperforms the fixed-length rule. This may be because, without an uncertainty threshold to inform stopping decisions, questions with higher uncertainty in answers might not provide enough information to make predictions, while questions with lower uncertainty in answers may provide unnecessary information they do not need. The variable length with caps rule might be used to achieve a balanced solution between prediction accuracy and the number of skipped questions.

5.4.3 Study 3: Simulation of One-Year Data Collection: Question Skipping and Real-Time Predictions.—In study 3, we simulated year-long data collection of both daily-EMA (Dataset 1) and hourly-EMA (Dataset 2) using our method to mimic implementation in the real world. Again, we benchmarked the performance by the number of questions skipped and real-time prediction accuracy on unselected questions.

Based on the results of Study 2, the simulation used the variable length (without cap) stopping rule with an uncertainty threshold of 0.4. For the forecasting setting, we took an expanding window approach to iteratively accumulate training datasets, as illustrated in Figure 12 in the Appendix. We started by using a period of training data with full survey responses to train the model and tested the performance in the following month. Then we progressed to testing on the next month by incorporating the testing dataset of the last iteration into the new training dataset. The incorporated testing dataset from the last iteration would include both observed responses to selected questions and predicted responses to unselected questions based on observed questions. Again, the DAG structure was learned using all participants' training data and the idiographic NPT was learned from each individual's training data.

Figure 5(1) and 5(2) show the one-month ahead prediction performance over time for daily and hourly-EMA using five response options. We found that as the model was trained on more survey data across time, the performance of the model increased steadily. For example, for daily-EMA with five options in Figure 5(1), the mean skipping percentage increased from 34.2% to 51.0% (i.e., progressively shorter surveys) while the mean prediction accuracy remained stable above 60.0%. A similar phenomenon was observed on the trajectory of hourly-EMA. The mean prediction accuracy and mean skipping percentage across the year (see more analysis about initial training periods and measurement precision in Study 7 and 8) are shown in Table 3. Compared to daily-EMA surveys, hourly-EMA demonstrated both enhanced prediction accuracy and skipping percentage. The enhanced performance might be because hourly-EMA collected data with less recall errors given shorter recall periods and the models were trained on a larger number of surveys in hourly-EMA datasets (three times higher sampling frequency than daily-EMA). For daily-EMA, the best model could skip 33.8% of questions overall with an accuracy of 71.8%; For hourly-EMA, the best model could skip 38.5% of questions overall with an accuracy of 77.5% on unselected questions, using five options and three-month data in the initial training period.

5.4.4 Study 4: Simulation of One-Year Data Collection: Impact of Lengths of Initial Full-Survey Training Periods.—

This study explored how the method performance would differ by training models on different lengths of initial full-survey training periods. We experimented with three different lengths of the initial training period (1 mo, 2 mo, 3 mo). By prolonging the length of the initial training period, we observed a trade-off between the number of questions skipped and prediction accuracy (Table 3). The mean prediction accuracy increased from 63.7% to 71.8%, and the mean skipping percentage decreased from 42.9% to 33.8%. This shows a minor performance improvement with an initial training period beyond one month and that one-month of intensive survey data might suffice to initialize the model.

5.4.5 Study 5: Simulation of One-Year Data Collection: Impact of Measurement Precision.—

This study examined the impact of measurement precision on performance. The measurement precision indicates the number of response options for Likert-style EMA survey questions. Past affect recognition research has benchmarked the classification performance using datasets with two or three classes of emotional states (e.g., positive, negative, neutral) [52, 72]. To compare the performance with this prior work, we reduced the measurement precision from five response options to three. The three options combined the first and last two options from the original five options (0 = “Not at all” and “A little,” 1 = “Moderately,” 2 = “Quite a bit” and “Extremely”).

Shown in Table 3 is the impact of using three classes for the same simulation as described in Study 3 (Section 5.4.3). As expected, with lower measurement precision, both prediction accuracy and skipping percentage increased greatly. For daily-EMA, the mean skipping percentage increased from 33.8%−42.9% to 48.1%−53.3% and the mean prediction accuracy increased from 63.7%−71.8% to 80.1%−85.2%. For hourly-EMA, the mean skipping percentage increased from 38.5%−43.7% to 51.9%−54.1% and the mean prediction accuracy increased from 73.2%−77.5% to 85.7%−87.4%.

5.4.6 Study 6: Missingness (Skipped Questions) Imputation of One-Year Simulated Datasets.—This study assessed whether our method could result in datasets with a higher amount of information than the planned missing data design via random omission (i.e., matrix sampling [58]) (**H2**). We examined how imputation error differs between the datasets simulated by our method and by planned missing data design via random omission. For comparison, datasets collected by planned missing data design were simulated by randomly selecting six questions from each complete-questions survey (random6). The datasets simulated using our method followed the same setting as Study 3 but using the variable-length stopping rule with cap. The purpose of capping the survey length is to ensure a fair comparison, with the resulting datasets simulated using our method having the same (or higher) percentage of skipped questions as those simulated by random-question omission. Different datasets were simulated using the cap of three, four, five, and six questions from each survey (cap3–6). Random6 and cap6 datasets should have 50% of planned missingness in simulated datasets, and cap3–5 datasets should have about 58%−75% data missing in total. Imputation error was estimated using mean squared error (MSE), calculated as the average squared difference between imputed and ground-truth values.

To impute missing data, we used the *IterativeImputer* method implemented in the Python *scikit-learn* package [44]. This method has a multiple imputation mechanism similar to the R *MICE* package (Multivariate Imputation by Chained Equations) [67], but differs from it by returning a single imputation instead of multiple imputations. Specifically, it estimates missing values of a feature as a function of other features and iterates for each feature at each round. The process repeats for 20 imputation rounds, and the results of the final imputation round are returned. We applied the same imputation method to both datasets simulated by our method and random omission.

Multiple dependent t-tests for paired samples were performed to compare participants’ average imputation errors between simulated datasets using random missingness (random6) and simulated datasets using our method with different cap levels (cap3–6). The distributions of participants’ average imputation errors of datasets are shown in Figure 6 and the results of statistical tests are summarized in Table 4. We found that cap5 and cap6 had significantly lower mean imputation errors compared to random6 by 9% and 15%, respectively (both p<0.05). The effect sizes, as measured by Cohen's d, were $d = 0.14$ and $d = 0.28$. The overall mean imputation errors of cap3 and cap4 were slightly higher than random6 but were not found significantly different ($d = -0.01$ and d = −0.10, indicating negligible differences).

5.4.7 Study 7: Generalizability Test on Two Additional Datasets.—

This study assessed whether our method could generalize to other EMA studies with different study populations, study durations, prompt schemes, and question sets (H3). We replicated Study 3 and Study 6 on two additional EMA datasets (Dataset 3 and Dataset 4).

Unlike Dataset 1 and Dataset 2 that only have questions on affect and feelings, the original EMA study of Dataset 3 employed EMA surveys that included a combination of questions regarding mental health states and behaviors (social contact and COVID-19 related). To build the BN model, we followed the same procedures as in Section 5.2 for Dataset 1 and Dataset 2. The challenge lies in the limited training data available from a two-week study to model 17 construct variables. To avoid some construct variables being isolated from other variables, the DAG structure was learned using two weeks' data from all participants and then each model was personalized by estimating NPT parameters using the first week of each individual's data. We evaluated the performance of the model on the second week's survey data.

The results are summarized in Table 5. By varying the stopping threshold from 0.2 to 0.4, our method allowed us to skip on average 45.0%−71.7% of questions from each survey with a real-time prediction accuracy of 77.3%−81.2% for skipped affect constructs. Compared to the dataset simulated by randomly omitting nine out of 17 questions from each survey (random8), our method (setting the stopping threshold to 0.3) results in a dataset collected with a similar amount of data missingness but lower imputation errors (cap8) (Figure 7). The paired-sample t-test showed that the mean of participant's average imputation errors of cap8 dataset (Mean = 0.58, SD = 0.33) were lower than those in the random8 dataset (Mean = 0.79, SD = 0.29) by about 27%, $t(73) = -7.3$, $p < 0.001$. The effect size $d = 0.65$ indicated a medium effect. The selected questions were evenly split between mental health questions (51.5%) and behavior questions (48.5%). The mean of participant's average imputation errors of skipped mental health questions in the cap8 dataset (Mean = 0.30, SD = 0.40) was lower than those in the random8 dataset (Mean = 0.38, SD = 0.44). The difference, however, was not statistically significant, $t(70) = -1.6$, $p = 0.11$, with an effect size $d = 0.05$. The mean of participant's average imputation errors of skipped behavior questions in cap8 dataset (Mean = 0.68, SD = 1.71) was lower than those in the random8 dataset (Mean = 0.78, SD = 1.38) and the difference was also not statistically significant, $t(69) = -0.94$, $p = 0.35$, with an effect size $d = 0.01$.

The original EMA study of Dataset 4 employed a composite scale to assess multi-dimensional mood constructs in precision for couples with personality disorders. Instead of building prediction models on individual question items, we built prediction models on the high-level affect constructs. This allows us to omit the entire question sets for the skipped affect constructs. The model building followed the same procedures as in Section 5.2 for

Dataset 1 and Dataset 2, except that only one week of data were used to initialize the model, including the DAG structure learning and NPT parameter estimation.

The results are summarized in Table 6. By setting the stopping threshold to 0.2, the simulated data collection showed that we could skip, on average, 60.3% (SD = 16.1%) of high-level affect constructs from each survey; survey questions used five response options. After correcting the number of question items shared between affect constructs, the proposed method allowed us to skip, on average, 53.3% (SD = 11.7%) of questions from each survey with a prediction accuracy of 94.5% (SD = 6.5%) for skipped high-level affect constructs.

Compared to the dataset simulated by randomly omitting three out of six constructs from each survey (random3), our method results in a dataset collected with a similar amount of data missingness but lower imputation errors (cap3) (Figure 8). The paired-sample t-test showed that cap3 had a statistically significant lower mean of participant's average imputation errors than random3 by 52% ($t(216) = -11.1$, $p < 0.001$, $d = 0.66$; medium effect).

6 DISCUSSION

In this work, we propose a question-selection method to reduce user response burden by shortening the survey length per prompt. By strategically skipping questions whose answers are confidently predicted from observed information, we proactively reduce user burden, improving user engagement with longitudinal data collection, while minimizing overall information loss. We demonstrated the feasibility of using the question informativeness in question-selection decisions using self-reported data from four real-world EMA datasets. By simulating the question selection of each prompted survey, we evaluated the performance of EMA data collection using the proposed method. Performance was evaluated by the percentage of questions skipped and the resulting information loss.

Our results show that it is feasible to quantify the information gain of each survey question using prediction uncertainty and to use the information gain to support the question-selection process. For both our method and random-question selection for each prompted survey, the prediction uncertainty decreased and prediction accuracy increased as more questions were answered. The rate of uncertainty reduction and accuracy enhancement, however, were higher when using our method than when using random-question selection, indicating the effectiveness of using prediction uncertainty to select informative questions compared to baseline random omission. In machine learning studies, prediction uncertainty is used to optimize training sample selection to improve model performance [49], whereas here we introduce the technique to optimize question selection in prompt decisions to improve the efficiency of intensive longitudinal EMA data collection. Previous research [53] reduced survey length by relying on psychometrically calibrated item banks. Our method, which learns the associations between constructs directly from an individual's prior responses, can support self-report data collection as long as the survey instruments measure multiple interrelated constructs. We validated our method's applicability across different question sets, including those with a mix of questions on mental states and behaviors

(Dataset 3) and those with composite multi-item scales (Dataset 4). The results demonstrate our method's flexibility in modeling and supporting question selection across different types and levels of self-reported constructs. Moreover, our method does not change sampling schemes that are often specific to study objectives, which makes it a practical method of reducing user response burden in most EMA data collection situations.

By simulating question selection using real-world datasets, we evaluated the number of questions skipped and the information loss by comparing inferred answers to skipped questions with the ground-truth labels. The results of simulating data collection on the year-long daily and hourly EMA survey datasets (Dataset 1 and 2) show that our method could greatly reduce the survey length while maintaining good real-time prediction performance on unselected questions throughout a year. By using full-survey responses from the first month to initiate model training, our method could achieve a reduction in survey length by 33% to 43% (about 4 or 5 out of 12 questions) with prediction accuracy of 64% to 72% for skipped daily EMA of five options; For hourly-EMA of five options, the method could achieve a reduction in survey length by 39% to 44% (about 5 out of 12 questions) with prediction accuracy of 73% to 78%. When reducing the measurement precision from five options into three, the method could achieve a skipping rate of 48% to 53% (about 6 out of 12 questions) with an accuracy of 80% to 85% for daily EMA and achieve a skipping rate of 52% to 54% (about 6 out of 12 questions) with an accuracy of 85% to 87% for hourly EMA. In EMA studies only lasting two to three weeks (Dataset 3 and 4), our method could still skip on average more than 50% of questions from each survey and achieve prediction accuracies of 80.7% (Dataset 3) and 94.5% (Dataset 4) for skipped questions with five options. When the data collection extends over a longer period, the average skipping percentage may increase over time while the prediction accuracy remains stable, as shown in Study 3 with the year-long dataset (Dataset 2).

Our results suggest that researchers using our method might be able to cut a survey length in half while incurring only a small prediction error rate on the unselected questions. Applied to real-time monitoring, our method could help digital health applications to deliver more personalized, just-in-time adaptive interventions (JITAIs) in real time by simultaneously monitoring more aspects of behavior and mental states. The prediction accuracy for skipped questions is comparable to studies that predict affective and emotional states using complex deep learning models trained on multimodal wearable sensing data [52, 72]. A recent review [72] reported studies achieving 74%−92% accuracy in three-emotion recognition using physiological signals. Predicting affective state is difficult because it varies greatly across occasions. One reason our method may achieve good prediction accuracy is because it inherently simplifies the prediction task by requesting labeling for difficult-to-predict (uncertain in answer) constructs and selectively makes predictions for easy-to-predict constructs. Another reason might be because our method uses a model that builds on the association between self-reported constructs assessed at each moment. From a participant's prior answer history, the model learns the participant's response patterns under certain types of contexts (e.g., the participant usually feels less energetic when stressed out) that may stay stable over time. The consistency of response patterns was validated in extreme cases, as shown in Dataset 3, where more informative questions could still be inferred from the first week's response data, despite participants experiencing significant life changes and shifts in mental health and social behavior due to government COVID-19 policies in the second week of the study.

From the perspective of researchers that conduct analysis using EMA datasets collected, we evaluated the missingness imputation performance of the dataset collected using our method. Compared to the dataset with questions randomly omitted from each complete-questions survey, we found datasets generated with questions selected by our method produced statistically significantly lower imputation errors with the same number of selected questions (Dataset 3 and 4) and even fewer questions per survey (Dataset 2). The results show that with a similar number of questions skipped, our method may result in lower imputation errors than planned random missingness or enable skipping of more questions with similar imputation errors. These results suggest that researchers using the method may cut the participants' completion time by half for each prompted survey without sacrificing the overall data quality. The method may also enable researchers to include even more survey questions in future EMA studies without inducing extra burden on participants.

Finally, we experimented with different stopping rules and model settings to guide real-world implementation. Experimental results showed the variable-length stopping rule outperformed the fixed-number stopping rule. The findings suggest that allowing the model to make decisions based on the informativeness of EMA questions may lead to optimal outcomes. In cases where question space is limited, one can also use variable length with a cap to limit the maximum survey length while leaving some space for the model to make decisions. For example, microinteraction-based EMA [28] prompts only one question at a time, but many more times per hour or day than standard EMA. Our method may make methods such as microinteraction-based EMA more feasible by enabling selection of the most informative questions to prompt. Lastly, our results indicate that increasing the initial training duration of answering full-set surveys beyond one month did not yield the anticipated performance improvement. Simulation results on Dataset 3 and Dataset 4 indicate that the method remained effective even when the model was trained on just one to two weeks of response data. The recent data may be more informative of the current states of people than data from many months prior, indicating that training the prediction model using the most recent data might be sufficient.

7 LIMITATIONS AND FUTURE RESEARCH

There are several limitations of this work that provide opportunities for future research. In this work, we evaluated our method by simulating EMA data collection based on real-world EMA datasets already collected. By simulating the omission of a subset of questions in each prompted survey, we evaluated the information loss by comparing predicted survey responses against ground truth survey responses obtained from participants. Previous researchers have demonstrated significant benefits of reducing survey length on enhancing participant compliance and mitigating perceived burden [14, 37]. From that work, it follows that if our proposed method can reduce the number of questions that must be asked at most prompts, it is likely to reduce perceived burden. Future work could include real-world studies to assess how surveys with varying number of questions and randomized question presentation order (a side-effect of our method) impact participant compliance and perceived burden compared to surveys with fixed number and order of questions. Although we show that our method results in datasets with lower imputation errors in post-study analyses compared with the planned missing data design [58], future work might include evaluation of the quality of the resulting datasets in statistical modeling tasks to examine the potential bias and standard errors in parameter estimation.

The naïve version of BN models we implemented could be extended in future work. In addition to mental state constructs with discrete response options, the prediction model can also be built on question-answer variables with continuous response values (e.g., when using visual analogue scales) by using discretization techniques [7] or hybrid Bayesian networks [51]. We demonstrated in this work that behavioral and mental constructs measured in self-report surveys could be modeled simultaneously for a comprehensive evaluation of a person. With advancements in wearable sensing [26], behavioral and contextual constructs may be derived from sensing data passively collected on wearable devices [77] and incorporated into the model to inform self-reported constructs. The prediction model we used in this work (Bayesian networks) only considers the associations between questions within any given single prompted survey. Although momentary mental states such as affect are transient in nature, recent research found emotional states at a given time point may be carried over to subsequent time points [32]. Future research should consider leveraging temporal associations between prompted surveys using dynamic models (e.g., dynamic Bayesian networks [38]).

Moreover, the data-driven approach of learning DAG structure requires a large amount of training data and may not be feasible in studies that both only last for weeks and use large and complex question sets. Future researchers may explore incorporating expert knowledge in structure estimation (e.g., setting soft and hard constraints in structure learning [30]) to reduce the amount of initial training data required to make accurate predictions or to avoid the cold start problem when adding new questions in surveys [2].

Finally, future researchers might investigate adjusting the length of surveys based on user context to further reduce response burden at inconvenient moments. Previous research, as summarized in Section 2.2, found EMA participants considered interruptions in specific contexts to be more disruptive to their daily lives than in other situations (e.g., during activities versus during activity transitions). Lengthy surveys presented at disruptive moments can be particularly burdensome. Therefore, shorter surveys at disruptive moments may reduce response burden more as compared to other moments. Given previous studies have successfully inferred times when people might be less receptive to EMAs using multimodal passive mobile sensing data (e.g., [36]), strategies of combining both question-answer information gain and momentary interruptibility in the prompt question selection process should be explored. For example, researchers could add a question in randomly selected surveys to collect self-report feedback on the disruptiveness of prompted surveys [71]. The alternative could be using non-response information, assuming participants may have a low probability of answering prompts in some moments [45]. With that additional information, momentary interruptibility may be predicted using commonly-measured passive sensing data (e.g., day of week, time of day, phone usage, location type). The estimated momentary interruptibility might then be used to adjust the stopping rule, thereby limiting the survey length according to the user's current capacity to be burdened.

8 CONCLUSION

User burden is one of the primary limitations of EMA in collecting high-quality intensive longitudinal data. In this work, we propose a machine-learning-based prompt-question-selection method to shorten the survey length for each prompt. Unlike prior work that changed sampling schemes or used psychometrically calibrated collections of questions, the method explicitly quantifies the potential question-answer information gain by modeling associations between surveyed constructs using prior survey response history. The results show that the method could reduce survey length with less information loss compared to random-question omission; the method might generalize to data collection tasks with different study objectives, populations, study durations, prompt schemes, and types of question sets. The proposed method might be used to create survey space for measuring more constructs and reduce response burden to improve user engagement in longitudinal data collection. For real-time intervention, this method might facilitate the real-time monitoring of various aspects of human behavior and states. Future research could explore deriving constructs from passive sensing data, leveraging temporal associations between survey questions, and combining question-answer information gain with momentary interruptibility in the question selection process.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health: NIH/NHLBI U01HL146327 and NIH/NCI R01CA240713. The content is solely the responsibility of the authors and does not represent the official views of the NIH. We thank our collaborators and the anonymous reviewers for their valuable feedback, which greatly improved this paper. We are especially grateful to Donald Hedeker for his guidance on the statistical analysis and to Sy-Miin Chow for her assistance in identifying open-source EMA datasets. We also acknowledge Krit Goyal, Adeeb M. Arif, Binod Thapa Chhetry, and Qu Tang for their contributions to the development of the TIME app and the associated data processing algorithms.

A APPENDICES

Table 7.

Twelve affect and feeling questions in daily and hourly-EMA surveys (Dataset 1 and 2). The 5-point response options include "Not at all," "A little," "Moderately," "Quite a bit," and "Extremely" or "Very much so."

Construct	Daily-EMA Question	Hourly-EMA Question
Happy	Over the past day, how HAPPY did you feel?	Right now, how HAPPY do you feel?
Energetic	Over the past day, how ENERGETIC did you feel?	Right now, how ENERGETIC do you feel?
Relaxed	Over the past day, how RELAXED did you feel?	Right now, how RELAXED do you feel?
Sad	Over the past day, how SAD did you feel?	Right now, how SAD do you feel?
Fatigue	Over the past day, how FATIGUED did you feel?	Right now, how FATIGUED do you feel?
Tense	Over the past day, how TENSE did you feel?	Right now, how TENSE do you feel?
Stressed	Over the past day, how STRESSED did you feel?	Right now, how STRESSED do you feel?
Frustrated	Over the past day, how FRUSTRATED did you feel?	Right now, how FRUSTRATED do you feel?

Construct	Daily-EMA Question	Hourly-EMA Question
Nervous	Over the past day, how NERVOUS did you feel?	Right now, how NERVOUS do you feel?
Attention	Over the past day, I felt FOCUSED.	Right now, I feel FOCUSED.
Self-control	Over the past day, I felt like I could RESIST doing things that aren't good for me.	Right now, I feel IN CONTROL.
Productivity	Over the past day, I PROCRASTINATED.	Right now, I am PROCRASTINATING.

Table 8.

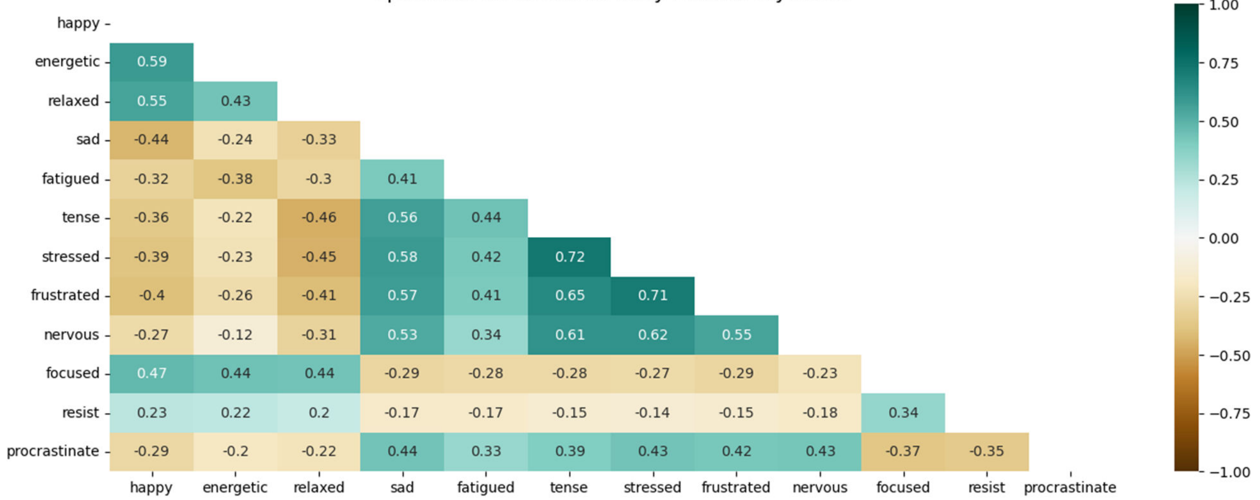
Eighteen mental state and behavior questions EMA surveys (Dataset 3). The 5-point response options include "Very slightly or not at all," "A little," "Moderately," "Quite a bit," and "Extremely" for mental health questions. The 5-point response options include "0 min," "1–15 min," "15–60 min," "1–2 hours," and "> 2 hours" for social contact and COVID-19-related behavior questions.

Construct	Question
Mental health (10)	
Relax	I found it difficult to relax.
Irritable	I felt (very) irritable.
Worry	I was worried about different things.
Nervous	I felt nervous, anxious or on edge.
Future	I felt that I had nothing to look forward.
Anhedonia	I couldn't seem to experience any positive feeling at all.
Tired	I felt tired.
Hunger	I was hungry.
Alone	I felt like I lack companionship, or that I am not close to people.
Angry	I felt angry.
Social contact behavior (5)	
Social offline	I spent ___ minutes on meaningful, offline, social interaction
Social online	I spent __ minutes using social media to kill/pass the time
Music	I spent __ minutes listening to music
Procrastination	To what degree did you postpone working on a task?
Time spent outdoors	I spent __ minutes outside (outdoors)?
COVID-19-related behavior (3)	
COVID-19 occupied	I spent __ occupied with the coronavirus (e.g. watching news thinking about it talking to friends about it)
COVID-19 worry	I spent __ thinking about my own health or that of my close friends and family members regarding the coronavirus
Home	I spent __ at home (including the home of parents/partner)

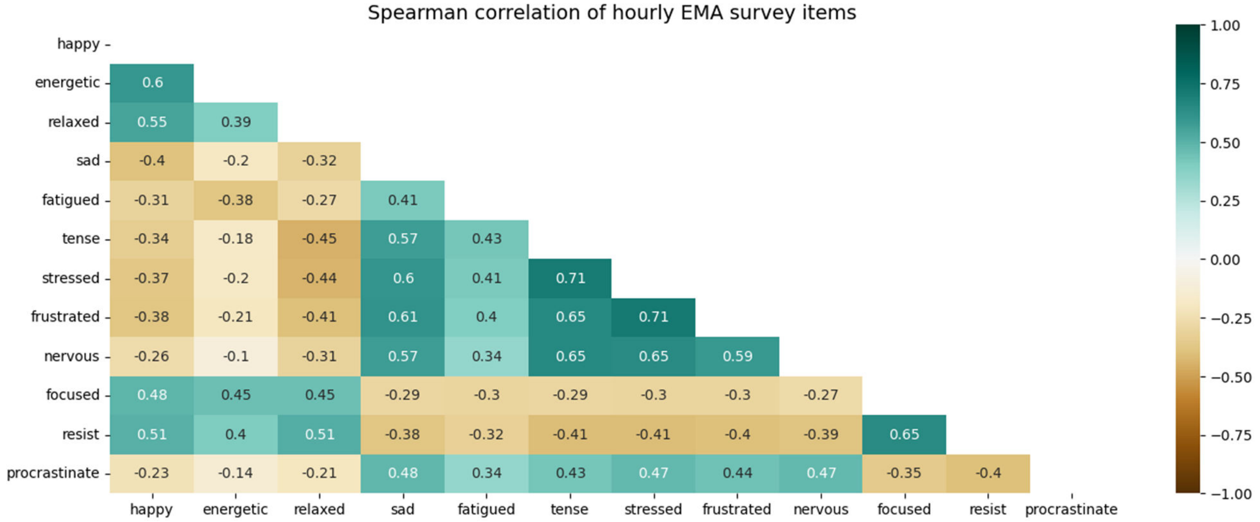
Table 9.

Item construction of six multi-item rating scales in event-contingent EMA surveys (Dataset 4). The 5-point response options include “Very slightly or not at all,” “A little,” “Moderately,” “Quite a bit,” and “Extremely.”

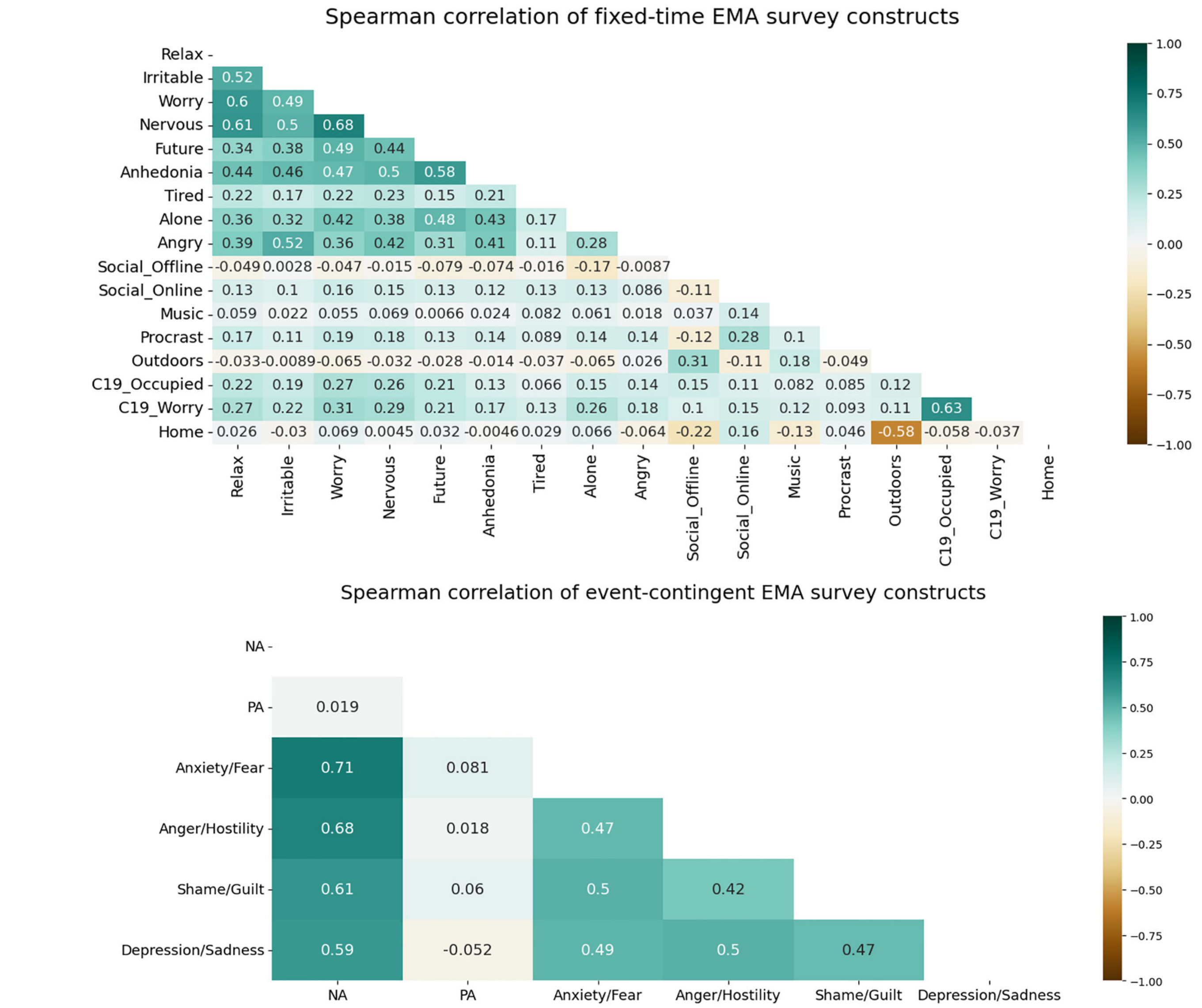
Construct	Item
General dimension scales	
Positive affect (10)	active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong
Negative affect (10)	afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed
Specific negative affect scales	
Fear (6)	afraid, scared, frightened, nervous, jittery, shaky
Hostility (6)	angry, hostile, irritable, scornful, disgusted, loathing
Guilt (2)	guilty, ashamed
Sadness (5)	sad, blue, downhearted, alone, lonely


Spearman correlation of daily EMA survey items

	happy	energetic	relaxed	sad	fatigued	tense	stressed	frustrated	nervous	focused	resist	procrastinate
happy												
energetic	0.59											
relaxed	0.55	0.43										
sad	-0.44	-0.24	-0.33									
fatigued	-0.32	-0.38	-0.3	0.41								
tense	-0.36	-0.22	-0.46	0.56	0.44							
stressed	-0.39	-0.23	-0.45	0.58	0.42	0.72						
frustrated	-0.4	-0.26	-0.41	0.57	0.41	0.65	0.71					
nervous	-0.27	-0.12	-0.31	0.53	0.34	0.61	0.62	0.55				
focused	0.47	0.44	0.44	-0.29	-0.28	-0.28	-0.27	-0.29	-0.23			
resist	0.23	0.22	0.2	-0.17	-0.17	-0.15	-0.14	-0.15	-0.18	0.34		
procrastinate	-0.29	-0.2	-0.22	0.44	0.33	0.39	0.43	0.42	0.43	-0.37	-0.35	


Spearman correlation of hourly EMA survey items

	happy	energetic	relaxed	sad	fatigued	tense	stressed	frustrated	nervous	focused	resist	procrastinate
happy												
energetic	0.6											
relaxed	0.55	0.39										
sad	-0.4	-0.2	-0.32									
fatigued	-0.31	-0.38	-0.27	0.41								
tense	-0.34	-0.18	-0.45	0.57	0.43							
stressed	-0.37	-0.2	-0.44	0.6	0.41	0.71						
frustrated	-0.38	-0.21	-0.41	0.61	0.4	0.65	0.71					
nervous	-0.26	-0.1	-0.31	0.57	0.34	0.65	0.65	0.59				
focused	0.48	0.45	0.45	-0.29	-0.3	-0.29	-0.3	-0.3	-0.27			
resist	0.51	0.4	0.51	-0.38	-0.32	-0.41	-0.41	-0.4	-0.39	0.65		
procrastinate	-0.23	-0.14	-0.21	0.48	0.34	0.43	0.47	0.44	0.47	-0.35	-0.4	



Spearman correlation of fixed-time EMA survey constructs

	Relax	Irritable	Worry	Nervous	Future	Anhedonia	Tired	Alone	Angry	Social_Offline	Social_Online	Music	Procrast	Outdoors	C19_Occupied	C19_Worry	Home
Relax																	
Irritable	0.52																
Worry	0.6	0.49															
Nervous	0.61	0.5	0.68														
Future	0.34	0.38	0.49	0.44													
Anhedonia	0.44	0.46	0.47	0.5	0.58												
Tired	0.22	0.17	0.22	0.23	0.15	0.21											
Alone	0.36	0.32	0.42	0.38	0.48	0.43	0.17										
Angry	0.39	0.52	0.36	0.42	0.31	0.41	0.11	0.28									
Social_Offline	-0.049	0.0028	-0.047	-0.015	-0.079	-0.074	-0.016	-0.17	-0.0087								
Social_Online	0.13	0.1	0.16	0.15	0.13	0.12	0.13	0.13	0.086	-0.11							
Music	0.059	0.022	0.055	0.069	0.0066	0.024	0.082	0.061	0.018	0.037	0.14						
Procrast	0.17	0.11	0.19	0.18	0.13	0.14	0.089	0.14	0.14	-0.12	0.28	0.1					
Outdoors	-0.033	-0.0089	-0.065	-0.032	-0.028	-0.014	-0.037	-0.065	0.026	0.31	-0.11	0.18	-0.049				
C19_Occupied	0.22	0.19	0.27	0.26	0.21	0.13	0.066	0.15	0.14	0.15	0.11	0.082	0.085	0.12			
C19_Worry	0.27	0.22	0.31	0.29	0.21	0.17	0.13	0.26	0.18	0.1	0.15	0.12	0.093	0.11	0.63		
Home	0.026	-0.03	0.069	0.0045	0.032	-0.0046	0.029	0.066	-0.064	-0.22	0.16	-0.13	0.046	-0.58	-0.058	-0.037	

Spearman correlation of event-contingent EMA survey constructs

	NA	PA	Anxiety/Fear	Anger/Hostility	Shame/Guilt	Depression/Sadness
NA						
PA	0.019					
Anxiety/Fear	0.71	0.081				
Anger/Hostility	0.68	0.018	0.47			
Shame/Guilt	0.61	0.06	0.5	0.42		
Depression/Sadness	0.59	-0.052	0.49	0.5	0.47	

Figure 9.

Correlation (Spearman) matrices of survey questions/constructs in daily (top) and hourly (bottom) EMA question sets.

Correlation (Spearman) matrices of survey questions/constructs in fixed-time (top) and event-contingent (bottom) EMA question sets.

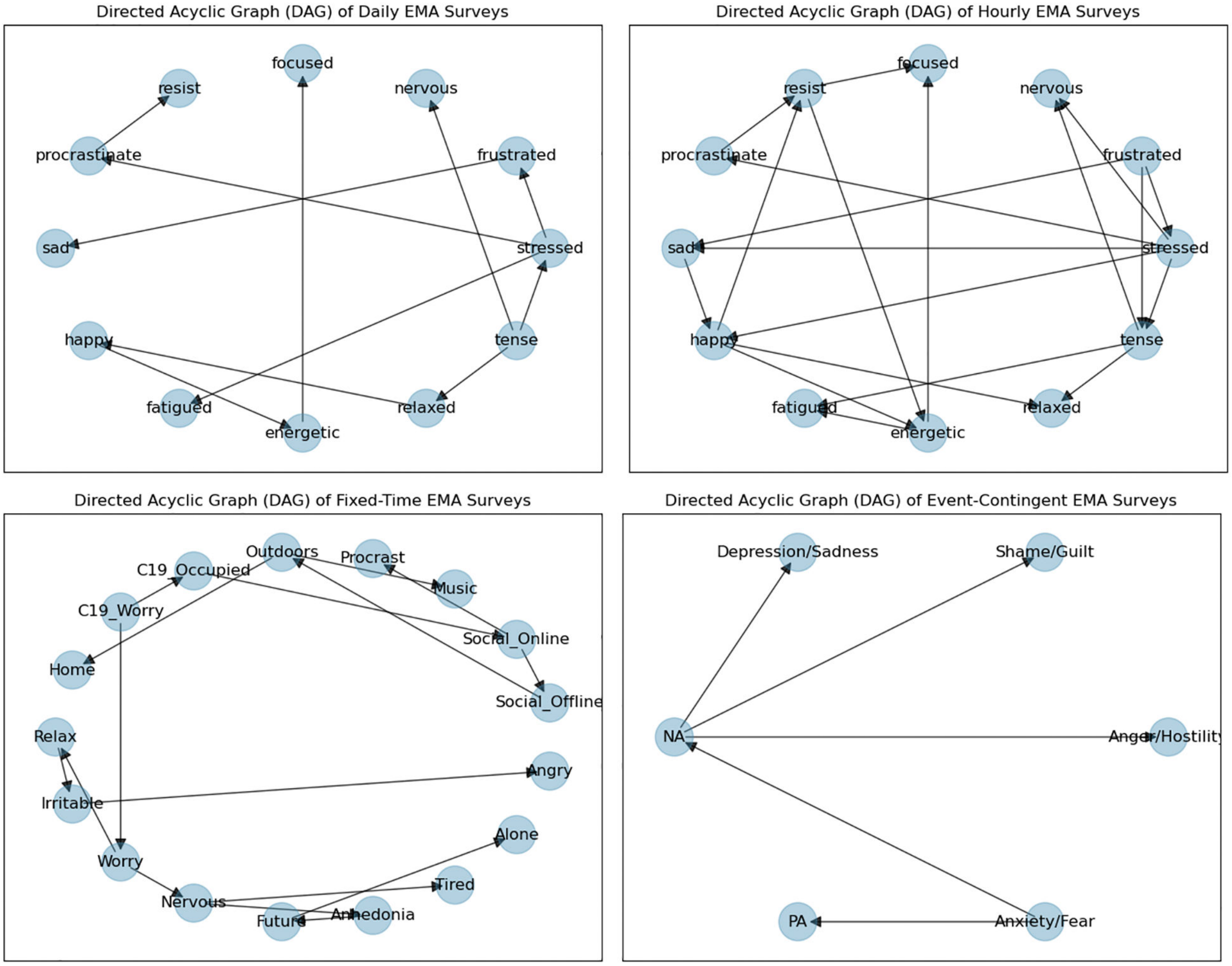


Figure 10.

DACs for daily EMA (top left), hourly EMA (top right), fixed-time EMA (bottom left), and event-contingent EMA (bottom right) surveys using initial training data (one month/one month/one week/one week) of all participants.

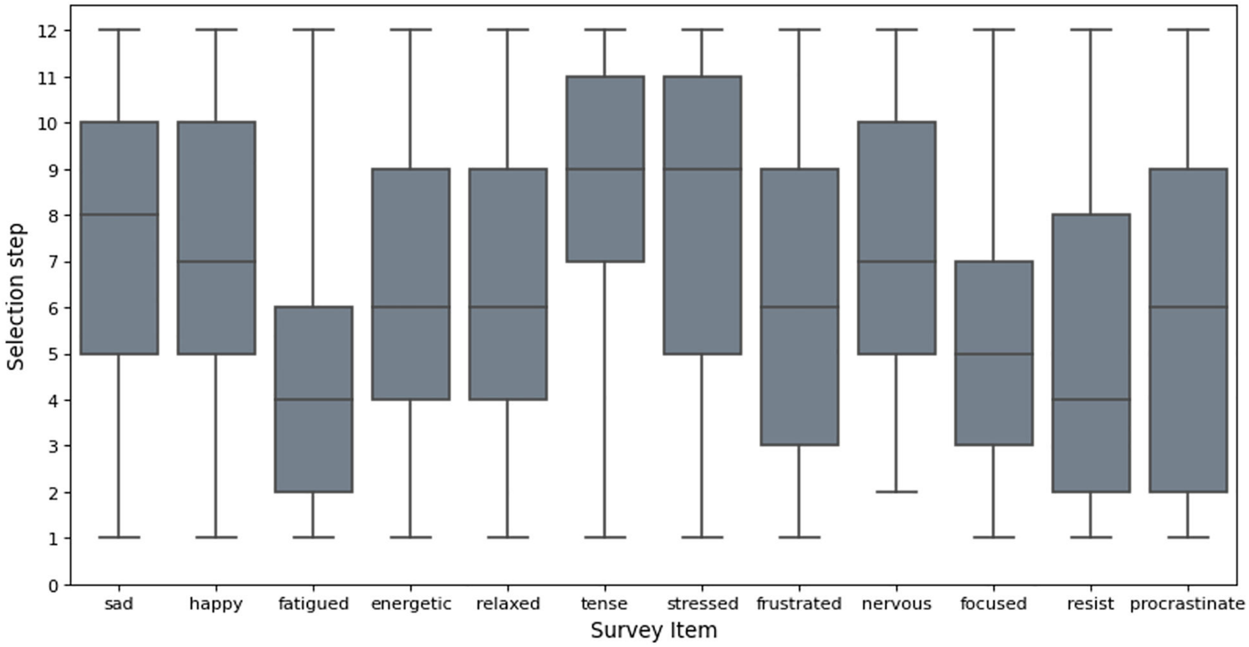


Figure 11.

Order number in which the questions were presented, where 12 indicates the question was presented last, and one indicates the question was presented first.

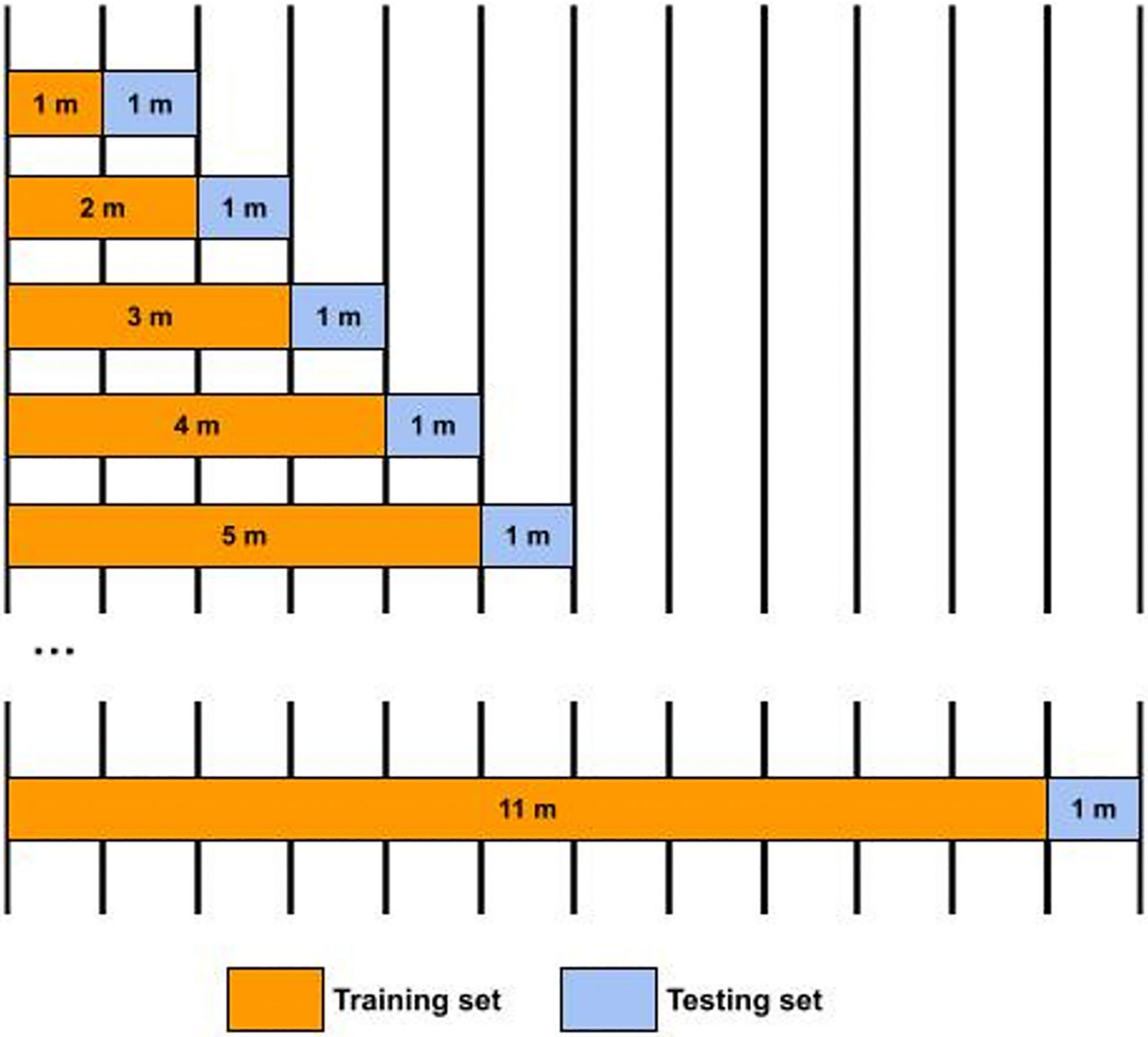


Figure 12.

Expanding window design of one-month ahead prediction across a year.

REFERENCES

[1]. Aminikhanghahi Samaneh, Schmitter-Edgecombe Maureen, and Cook Diane J. 2019. Context-aware delivery of ecological momentary assessment. IEEE Journal of Biomedical and Health Informatics, 24(4): p. 1206–1214. [PubMed: 31443058]

[2]. Amirkhani Hossein, Rahmati Mohammad, Lucas Peter JF, and Hommersom Arjen. 2016. Exploiting experts' knowledge for structure learning of Bayesian networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): p. 2154–2170. [PubMed: 28114005]

[3]. Balaskas Andreas, Schueller Stephen M, Cox Anna L, and Doherty Gavin. 2021. Ecological momentary interventions for mental health: A scoping review. PLOS One, 16(3): p. e0248152. [PubMed: 33705457]

[4]. Bayes Thomas. 1991. An essay towards solving a problem in the doctrine of chances. 1763. MD Computing: Computers in Medical Practice, 8(3): p. 157–171. [PubMed: 1857193]

[5]. Berkel Niels van, Luo Chu, Anagnostopoulos Theodoros, Ferreira Denzil, Goncalves Jorge, Hosio Simo, and Kostakos Vassilis. 2016. A Systematic assessment of smartphone usage gaps, in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery. p. 4711–4721.

[6]. Chen Chen, Lifset Ella T, Han Yichen, Roy Arkajyoti, Hogarth Michael, Moore Alison A, Farcas Emilia, and Weibel Nadir. 2023. Screen or no screen? Lessons learnt from a real-world deployment study of using voice assistants with and without touchscreen for older adults. In Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility. p. 1–21.

[7]. Chen Yi-Chun, Wheeler Tim A, and Kochenderfer Mykel J. 2017. Learning discrete Bayesian networks from continuous data. Journal of Artificial Intelligence Research, 59: p. 103–132.

[8]. Chevance Guillaume, Golaszewski Natalie M, Baretta Dario, Hekler Eric B, Larsen Britta A, Patrick Kevin, and Godino Job. 2020. Modelling multiple health behavior change with network analyses: Results from a one-year study conducted among overweight and obese adults. Journal of Behavioral Medicine, 43: p. 254–261. [PubMed: 31997127]

[9]. Consolvo S and Walker M. 2003. Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Computing, 2(2): p. 24–31.

[10]. Csikszentmihalyi M and Larson R. 1987. Validity and reliability of the experience-sampling method. The Journal of Nervous and Mental Disease, 175(9): p. 526–36. [PubMed: 3655778]

[11]. Curran ShellyL, Beacham AbbieO, and Andrykowski MichaelA. 2004. Ecological momentary assessment of fatigue following breast cancer treatment. Journal of Behavioral Medicine, 27(5): p. 425–444. [PubMed: 15675633]

[12]. Dunton Genevieve F, Rothman Alexander J, Leventhal Adam M, and Intille Stephen S. 2021. How intensive longitudinal data can stimulate advances in health behavior maintenance theories and interventions. Translational Behavioral Medicine, 11(1): p. 281–286. [PubMed: 31731290]

[13]. Dzubur Eldin. 2017. Understanding the methodological limitations in the ecological momentary assessment of physical activity. PhD Thesis, Department, University of Southern California.

[14]. Eisele Gudrun, Vachon Hugo, Lafit Ginette, Kuppens Peter, Houben Marlies, Myin-Germeys Inez, and Viechtbauer Wolfgang. 2022. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment, 29(2): p. 136–151. [PubMed: 32909448]

[15]. Fenton Norman and Neil Martin, Risk Assessment and Decision Analysis with Bayesian Networks. Second ed. 2018: CRC Press.

[16]. Ferreira Denzil, Goncalves Jorge, Kostakos Vassilis, Barkhuus Louise, and Dey Anind K., Contextual experience sampling of mobile application micro-usage, in Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services. 2014, ACM: Toronto, ON, Canada. p. 91–100.

[17]. Fogarty James, Ko Andrew J., Aung Htet Htet, Golden Elspeth, Tang Karen P., and Hudson Scott E.. 2005. Examining task engagement in sensor-based statistical models of human interruptibility, in Proceedings of the SIGCHI Conference on Human factors in Computing Systems. ACM Press: Portland, Oregon, USA. p. 331–340.

[18]. Fried Eiko I, Papanikolaou Faidra, and Epskamp Sacha. 2022. Mental health and social contact during the COVID-19 pandemic: An ecological momentary assessment study. Clinical Psychological Science, 10(2): p. 340–354.

[19]. Fuller-Tyszkiewicz Matthew, Skouteris Helen, Richardson Ben, Blore Jed, Holmes Millicent, and Mills Jacqueline. 2013. Does the burden of the experience sampling method undermine data quality in state body image research? Body Image, 10(4): p. 607–13. [PubMed: 23856302]

[20]. Geiger Dan, Verma Thomas, and Pearl Judea. 1990. Identifying independence in Bayesian networks. Networks, 20(5): p. 507–534.

[21]. Gibbons Chris J. 2017. Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century. Frontiers in Psychology, 7: p. 1933. [PubMed: 28154540]

[22]. Gibbons Chris J.. 2016. Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st Century. Front Psychol, 7: p. 1933. [PubMed: 28154540]

[23]. Graham John W, Taylor Bonnie J, Olchowski Allison E, and Cumsille Patricio E. 2006. Planned missing data designs in psychological research. Psychological Methods, 11(4): p. 323. [PubMed: 17154750]

[24]. Han Yichen, Han Christopher Bo, Chen Chen, Lee Peng Wei, Hogarth Michael, Moore Alison A, Weibel Nadir, and Farcas Emilia. 2022. Towards visualization of time–series ecological momentary assessment (EMA) data on standalone voice–first virtual assistants. In Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22). Association for Computing Machinery, New York, NY, United States, p. 1–4.

[25]. Hedeker D and Nordgren R. 2013. MIXREGLS: A program for mixed-effects location scale analysis. Journal of Statistical Software, 52(12): p. 1–38. [PubMed: 23761062]

[26]. Hoemann Katie, Wormwood Jolie B, Barrett Lisa Feldman, and Quigley Karen S. 2023. Multimodal, idiographic ambulatory sensing will transform our understanding of emotion. Affective Science, 4(3): p. 480–486. [PubMed: 37744967]

[27]. Hufford Michael R., Shields Alan L., Shiffman Saul, Paty Jean, and Balabanis Mark. 2002. Reactivity to ecological momentary assessment: An example using undergraduate problem drinkers. Psychology of Addictive Behaviors, 16(3): p. 205–211. [PubMed: 12236455]

[28]. Intille Stephen, Haynes Caitlin, Maniar Dharam, Ponnada Aditya, and Manjourides Justin. 2016. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16). Association for Computing Machinery, New York, NY, United States, p. 1124–1128.

[29]. Kim Jinhyuk, Marcusson-Clavertz David, Yoshiuchi Kazuhiro, and Smyth Joshua M.. 2019. Potential benefits of integrating ecological momentary assessment data into mHealth care systems. BioPsychoSocial Medicine, 13(1): p. 19. [PubMed: 31413726]

[30]. Kitson Neville Kenneth, Constantinou Anthony C, Guo Zhigao, Liu Yang, and Chobtham Kiattikun. 2023. A survey of Bayesian Network structure learning. Artificial Intelligence Review, 56(8): p. 8721–8814.

[31]. Klasnja Predrag, Hekler Eric B., Shiffman Saul, Boruvka Audrey, Almirall Daniel, Tewari Ambuj, and Murphy Susan A.. 2015. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association, 34S(0): p. 1220–1228. [PubMed: 26651463]

[32]. Koval Peter, Sütterlin Stefan, and Kuppens Peter. 2016. Emotional inertia is associated with lower well-being when controlling for differences in emotional context. Frontiers in Psychology, 6: p. 1997. [PubMed: 26779099]

[33]. Kyrimi Evangelia, McLachlan Scott, Dube Kudakwashe, Mariana R Neves Ali Fahmi, and Fenton Norman. 2021. A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. Artificial Intelligence in Medicine, 117: p. 102108. [PubMed: 34127238]

[34]. May Marcella, Junghaenel Doerte U, Ono Masakatsu, Stone Arthur A, and Schneider Stefan. 2018. Ecological momentary assessment methodology in chronic pain research: A systematic review. The Journal of Pain, 19(7): p. 699–716. [PubMed: 29371113]

[35]. Mehrotra A, Musolesi M, Hendley R, and Pejovic V, Designing content-driven intelligent notification mechanisms for mobile applications, in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015, ACM: Osaka, Japan. p. 813–824.

[36]. Mishra Varun, Lowens Byron, Lord Sarah, Caine Kelly, and Kotz David, Investigating contextual cues as indicators for EMA delivery, in Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. 2017, ACM: Maui, Hawaii. p. 935–940.

[37]. Morren M, van Dulmen S, Ouwerkerk J, and Bensing J. 2009. Compliance with momentary pain measurement using electronic diaries: A systematic review. European journal of pain, 13(4): p. 354–65. [PubMed: 18603458]

[38]. Murphy Kevin Patrick. 2002. Dynamic bayesian networks: representation, inference and learning. Thesis, Department, University of California, Berkeley.

[39]. Murray Aja Louise, Brown Ruth, Zhu Xinxin, Speyer Lydia Gabriela, Yang Yi, Xiao Zhouni, Ribeaud Denis, and Eisner Manuel. 2023. Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. Journal of Affective Disorders, 322: p. 125–131. [PubMed: 36372127]

[40]. Neath Andrew A and Cavanaugh Joseph E. 2012. The Bayesian information criterion: Background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2): p. 199–203.

[41]. Obuchi Mikio, Sasaki Wataru, Okoshi Tadashi, Nakazawa Jin, and Tokuda Hideyuki, Investigating interruptibility at activity breakpoints using smartphone activity recognition API, in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. 2016, ACM: Heidelberg, Germany. p. 1602–1607.

[42]. Ouali Yassine, Hudelot Céline, and Tami Myriam. 2020. An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278.

[43]. Park Chunjong, Lim Junsung, Kim Juho, Lee Sung-Ju, and Lee Dongman, Don't bother me. I'm socializing! A breakpoint-based smartphone notification system, in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). 2017, Association for Computing Machinery, New York, NY, United States: Portland, Oregon, USA. p. 541–554.

[44]. Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, and Dubourg Vincent. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: p. 2825–2830.

[45]. Ponnada Aditya, Li Jixin, Wang Shirlene, Wang Wei-Lin, Do Bridgette, Dunton Genevieve F., and Intille Stephen S.. 2022. Contextual biases in microinteraction ecological momentary assessment (μEMA) non-response. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(1): p. 1–24.

[46]. Ponnada Aditya, Wang Shirlene, Chu Daniel, Do Bridgette, Dunton Genevieve, and Intille Stephen. 2022. Intensive longitudinal data collection using microinteraction ecological momentary assessment: Pilot and preliminary results. JMIR Formative Research, 6(2): p. e32772. [PubMed: 35138253]

[47]. Ram Nilam, Brinberg Miriam, Pincus Aaron L, and Conroy David E. 2017. The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. Research in Human Development, 14(3): p. 253–270. [PubMed: 30613195]

[48]. Reininghaus Ulrich, Kempton Matthew J, Valmaggia Lucia, Craig Tom KJ, Garety Philippa, Onyejiaka Adanna, Gayer-Anderson Charlotte, So Suzanne H, Hubbard Kathryn, and Beards Stephanie. 2016. Stress sensitivity, aberrant salience, and threat anticipation in early psychosis: An experience sampling study. Schizophrenia Bulletin, 42(3): p. 712–722. [PubMed: 26834027]

[49]. Ren Pengzhen, Xiao Yun, Chang Xiaojun, Huang Po-Yao, Li Zhihui, Gupta Brij B, Chen Xiaojiang, and Wang Xin. 2021. A survey of deep active learning. ACM Computing Surveys (CSUR), 54(9): p. 1–40.

[50]. Russell Stuart J and Norvig Peter, Artificial intelligence a modern approach. 2010: London.

[51]. Salmerón Antonio, Rumí Rafael, Langseth Helge, Nielsen Thomas D, and Madsen Anders L. 2018. A review of inference algorithms for hybrid Bayesian networks. Journal of Artificial Intelligence Research, 62: p. 799–828.

[52]. Schmidt Philip, Reiss Attila, Dürichen Robert, and Laerhoven Kristof Van. 2019. Wearable-based affect recognition—A review. Sensors, 19(19): p. 4079. [PubMed: 31547220]

[53]. Schneider Stefan, Junghaenel Doerte U, Smyth Joshua M, Fred Wen Cheng K, and Stone Arthur A. 2024. Just-in-time adaptive ecological momentary assessment (JITA-EMA). Behavior Research Methods, 56: p. 765–783. [PubMed: 36840916]

[54]. Settles Burr. 2009. Active learning literature survey. Thesis, Department, University of Wisconsin-Madison.

[55]. Shannon Claude Elwood. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27(3): p. 379–423.

[56]. Shiffman Saul, Stone Arthur A., and Hufford Michael R.. 2008. Ecological momentary assessment. Annu Rev Clin Psychol, 4: p. 1–32. [PubMed: 18509902]

[57]. Shiyko MP, Burkhalter J, Li R, and Park BJ. 2014. Modeling nonlinear time-dependent treatment effects: An application of the generalized time-varying effect model (TVEM). J Consult Clin Psychol, 82(5): p. 760–72. [PubMed: 24364799]

[58]. Silvia Paul J, Kwapil Thomas R, Walsh Molly A, and Myin-Germeys Inez. 2014. Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. Behavior Research Methods, 46: p. 41–54. [PubMed: 23709167]

[59]. Smyth Joshua M, Jones Dusti R, Wen Cheng K F, Materia Frank T, Schneider Stefan, and Stone Arthur. 2021. Influence of ecological momentary assessment study design features on reported willingness to participate and perceptions of potential research studies: An experimental study. BMJ Open, 11(7): p. e049154.

[60]. Stone AA, Kessler RC, and Haythornthwaite JA. 1991. Measuring daily events and experiences: Decisions for the researcher. Journal of Personality, 59(3): p. 575–607. [PubMed: 1960643]

[61]. Stone Arthur A. and Shiffman Saul. 1994. Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine, 16(3): p. 199–202.

[62]. Stone Arthur A., Schneider Stefan, and Smyth Joshua M.. 2022. Evaluation of pressing issues in ecological momentary assessment. Annual Review of Clinical Psychology.

[63]. Taskesen Erdogan. 2020. Learning Bayesian Networks with the bnlearn Python Package. Retrieved October 11, 2024 from https://erdogant.github.io/bnlearn

[64]. Timms Kevin P, Rivera Daniel E, Collins Linda M, and Piper Megan E. 2013. A dynamical systems approach to understanding self-regulation in smoking cessation behavior change. Nicotine & Tobacco Research, 16(Suppl_2): p. S159–S168. [PubMed: 24064386]

[65]. van Berkel Niels, Ferreira Denzil, and Kostakos Vassilis. 2017. The experience sampling method on mobile devices. ACM Computing Surveys, 50(6): p. 1–40.

[66]. van Berkel Niels, Goncalves Jorge, Hosio Simo, Sarsenbayeva Zhanna, Velloso Eduardo, and Kostakos Vassilis. 2020. Overcoming compliance bias in self-report studies: A cross-study analysis. International Journal of Human-Computer Studies, 134: p. 1–12.

[67]. Van Buuren Stef and Groothuis-Oudshoorn Karin. 2011. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3): p. 1–67.

[68]. van Roekel Eeske, Keijsers Loes, and Chung Joanne M. 2019. A review of current ambulatory assessment studies in adolescent samples and practical recommendations. Journal of Research on Adolescence, 29(3): p. 560–577. [PubMed: 31573762]

[69]. Visuri Aku, Van Berkel Niels, Luo Chu, Goncalves Jorge, Ferreira Denzil, and Kostakos Vassilis. 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 (MobileHCI '17). Association for Computing Machinery, New York, NY, United States, p. 1–14.

[70]. Wang Shirlene. 2023. Developing and Testing Novel Strategies to Detect Inattentive Responding in Ecological Momentary Assessment Studies. PhD Thesis Thesis, Department, University of Southern California.

[71]. Wang Shirlene, Intille Stephen, Ponnada Aditya, Do Bridgette, Rothman Alexander, and Dunton Genevieve. 2022. Investigating microtemporal processes underlying health behavior adoption and maintenance: Protocol for an intensive longitudinal observational study. JMIR Research Protocols, 11(7): p. e36666. [PubMed: 35834296]

[72]. Wang Yan, Song Wei, Tao Wei, Liotta Antonio, Yang Dawei, Li Xinlei, Gao Shuyong, Sun Yixuan, Ge Weifeng, and Zhang Wei. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83: p. 19–52.

[73]. Watson David and Clark Lee Anna. 1994. The PANAS-X: Manual for the positive and negative affect scheduleexpanded form. Unpublished manuscript, University of Iowa.

[74]. Wenze Susan J and Miller Ivan W. 2010. Use of ecological momentary assessment in mood disorders research. Clinical Psychology Review, 30(6): p. 794–804. [PubMed: 20619520]

[75]. Wilhelm Peter and Schoebi Dominik. 2007. Assessing mood in daily life. European Journal of Psychological Assessment, 23(4): p. 258–267.

[76]. Wright Aidan GC, Stepp Stephanie D, Scott Lori N, Hallquist Michael N, Beeney Joseph E, Lazarus Sophie A, and Pilkonis Paul A. 2017. The effect of pathological narcissism on interpersonal and affective processes in social interactions. Journal of Abnormal Psychology, 126(7): p. 898. [PubMed: 29106275]

[77]. Yu Han and Sano Akane. 2023. Semi-supervised learning for wearable-based momentary stress detection in the wild. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(2): p. 1–23.

[78]. Zapata-Lamana Rafael, Lalanza Jaume F., Losilla Josep-Maria, Parrado Eva, and Capdevila Lluis. 2020. mHealth technology for ecological momentary assessment in physical activity research: A systematic review. PeerJ, 8: p. e8848–e8848. [PubMed: 32257648]

[79]. Zhang Xiaoyi, Pina Laura R., and Fogarty James, Examining unlock journaling with diaries and reminders for In situ self-report in health and wellness, in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16). 2016, Association for Computing Machinery, New York, NY, United States: Santa Clara, California, USA. p. 5658–5664.

CCS CONCEPTS

•Applied computing~Life and medical sciences~Health informatics•Human-centered computing~Human computer interaction (HCI)~Interactive systems and tools•Computing methodologies~Machine learning

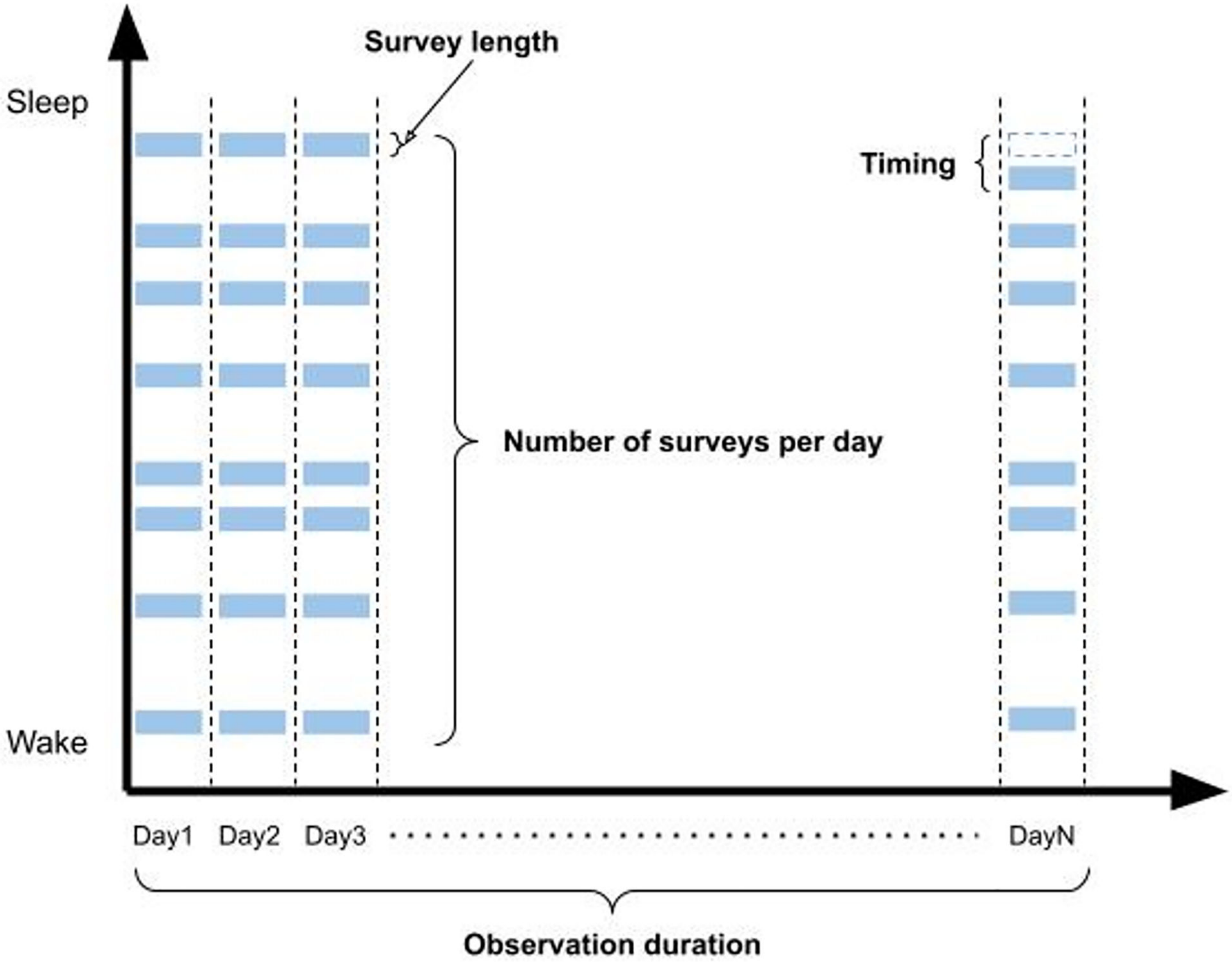
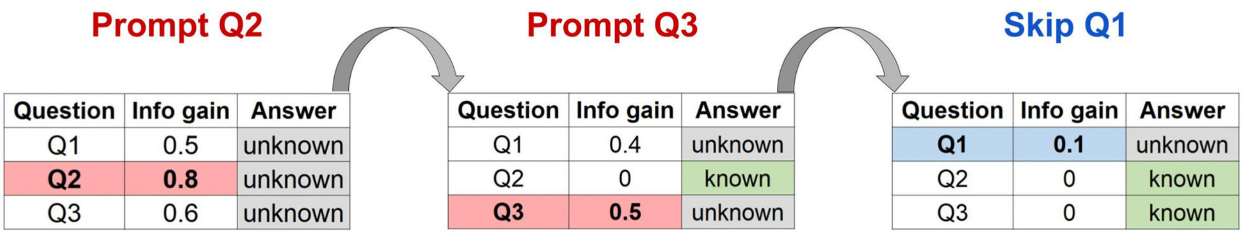


Figure 1.

Timing-related and data-related design components of EMA for user burden reduction. Each blue rectangle represents a prompted survey, with the size of the rectangle indicating the survey length. Survey length, number of surveys per day, observation duration, and prompt timing are the key EMA design components that can be adjusted by algorithms to reduce user response burden.



Prompt Q2

Question	Info gain	Answer
Q1	0.5	unknown
Q2	**0.8**	unknown
Q3	0.6	unknown

Prompt Q3

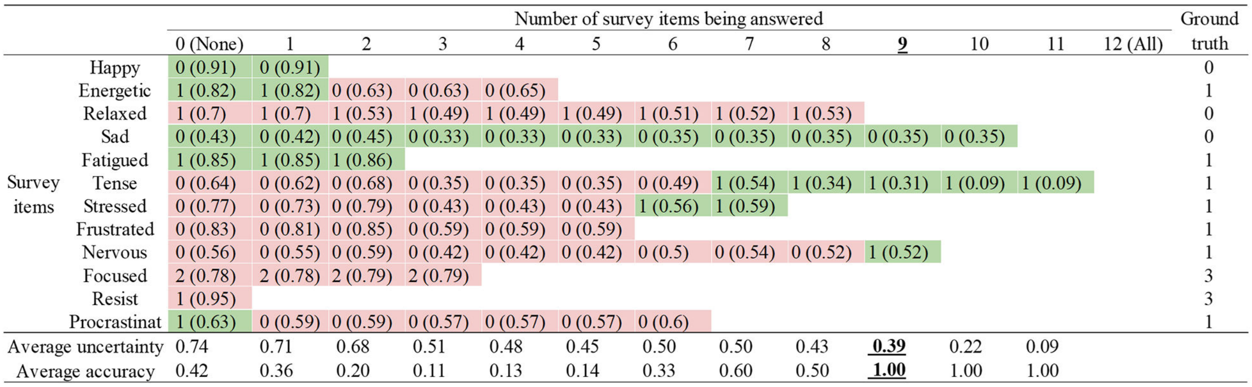
Question	Info gain	Answer
Q1	0.4	unknown
Q2	0	known
Q3	**0.5**	unknown

Skip Q1

Question	Info gain	Answer
Q1	**0.1**	unknown
Q2	0	known
Q3	0	known

Figure 2.

Conceptual illustration of the question selection process of a hypothesized EMA survey. The survey contains three questions. For each step, the EMA system selects the most informative question to be presented. After observing the responses to Q2 and Q3 in sequence, the system finds the information gain of Q1 is so low that Q1 can be confidently skipped with a high prediction accuracy. As a result, the length of this prompted survey is reduced from three to two with minimal information loss.



		Number of survey items being answered													Ground
		0 (None)	1	2	3	4	5	6	7	8	**<u>9</u>**	10	11	12 (All)	truth
	Happy	🟩 0 (0.91)	🟩 0 (0.91)												0
	Energetic	🟩 1 (0.82)	🟩 1 (0.82)	🟥 0 (0.63)	🟥 0 (0.63)	🟥 0 (0.65)									1
	Relaxed	🟥 1 (0.7)	🟥 1 (0.7)	🟥 1 (0.53)	🟥 1 (0.49)	🟥 1 (0.49)	🟥 1 (0.49)	🟥 1 (0.51)	🟥 1 (0.52)	🟥 1 (0.53)					0
	Sad	🟩 0 (0.43)	🟩 0 (0.42)	🟩 0 (0.45)	🟩 0 (0.33)	🟩 0 (0.33)	🟩 0 (0.33)	🟩 0 (0.35)	🟩 0 (0.35)	🟩 0 (0.35)	🟩 0 (0.35)	🟩 0 (0.35)			0
	Fatigued	🟩 1 (0.85)	🟩 1 (0.85)	🟩 1 (0.86)											1
Survey	Tense	🟥 0 (0.64)	🟥 0 (0.62)	🟥 0 (0.68)	🟥 0 (0.35)	🟥 0 (0.35)	🟥 0 (0.35)	🟥 0 (0.49)	🟩 1 (0.54)	🟩 1 (0.34)	🟩 1 (0.31)	🟩 1 (0.09)	🟩 1 (0.09)		1
items	Stressed	🟥 0 (0.77)	🟥 0 (0.73)	🟥 0 (0.79)	🟥 0 (0.43)	🟥 0 (0.43)	🟥 0 (0.43)	🟩 1 (0.56)	🟩 1 (0.59)						1
	Frustrated	🟥 0 (0.83)	🟥 0 (0.81)	🟥 0 (0.85)	🟥 0 (0.59)	🟥 0 (0.59)	🟥 0 (0.59)								1
	Nervous	🟥 0 (0.56)	🟥 0 (0.55)	🟥 0 (0.59)	🟥 0 (0.42)	🟥 0 (0.42)	🟥 0 (0.42)	🟥 0 (0.5)	🟥 0 (0.54)	🟥 0 (0.52)	🟩 1 (0.52)				1
	Focused	🟥 2 (0.78)	🟥 2 (0.78)	🟥 2 (0.79)	🟥 2 (0.79)										3
	Resist	🟥 1 (0.95)													3
	Procrastinat	🟩 1 (0.63)	🟥 0 (0.59)	🟥 0 (0.59)	🟥 0 (0.57)	🟥 0 (0.57)	🟥 0 (0.57)	🟥 0 (0.6)							1
Average uncertainty		0.74	0.71	0.68	0.51	0.48	0.45	0.50	0.50	0.43	**<u>0.39</u>**	0.22	0.09		
Average accuracy		0.42	0.36	0.20	0.11	0.13	0.14	0.33	0.60	0.50	**<u>1.00</u>**	1.00	1.00		

Figure 3.

Breakdown of prompt-question selection process of one daily-EMA survey. For each step of question selection, one survey question with the highest uncertainty was picked and the response was obtained from the participant. Each column represents a snapshot of predicted responses and prediction uncertainty of unselected survey question items after question selection of each step. Cell values in the survey question rows indicate predicted responses (integer 0–4) and prediction uncertainty (float 0–1, in parenthesis) for the corresponding unselected question of the row. The last column shows ground-truth responses to all questions. The cell color is green if the predicted response at that step matches the ground-truth responses; otherwise, it is red. The bottom two rows are the average prediction uncertainty and accuracy of all unselected questions for each step. Note that the optimal strategy is to stop asking questions after nine questions get answered by the participant, cutting the survey short for three questions (sad, tense, and frustrated), which are correctly predicted by the model.

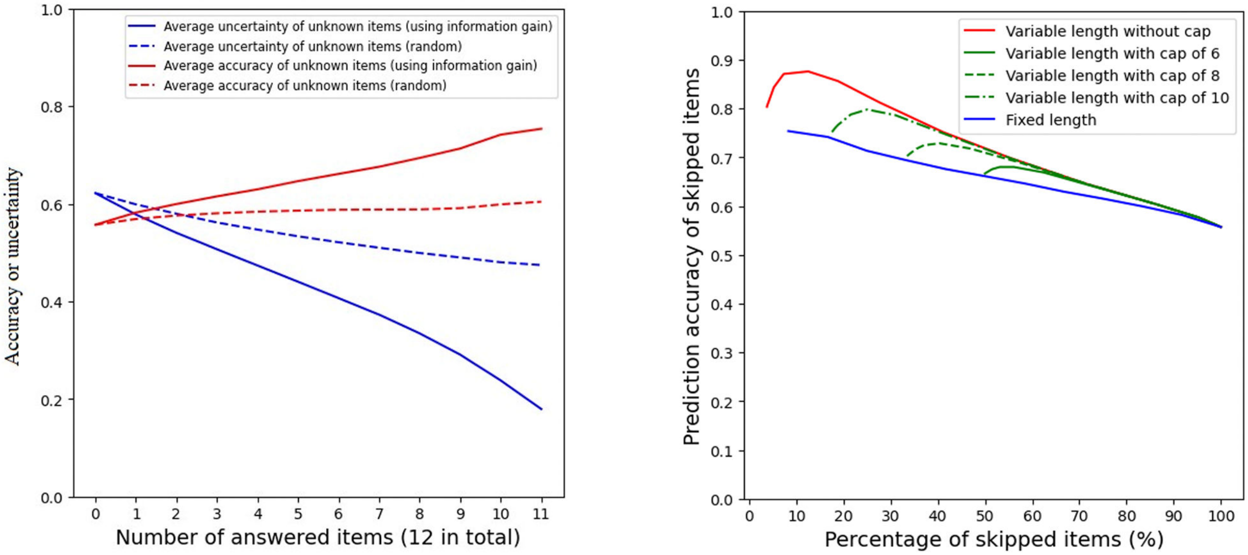


Figure 4.

(left) Question selection using information gain versus at random. The slopes of reduction in uncertainty level of unknown questions and increase in prediction accuracy are much steeper when using information gain in question selection, (right) Comparison of stopping rules: fixed length, variable length, and variable length with cap. Different lengths (fixed length) and thresholds (variable length) changed the trade-off between skipping percentage and prediction accuracy. Overall, the stopping rule of variable length without cap outperforms variable length with cap, followed by the fixed-length rule.

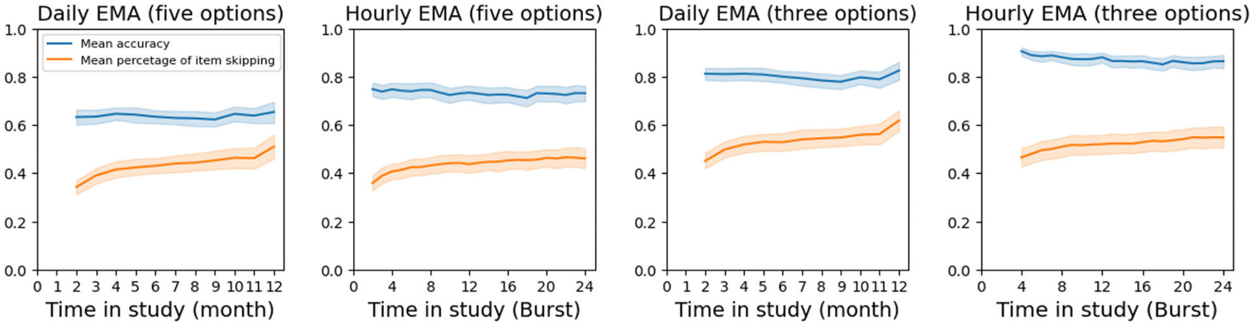


Figure 5.

One-month ahead prediction performance of simulated data collection for (1) daily EMA with five options, (2) hourly EMA with five options, (3) daily EMA with three options, and (4) hourly EMA with three options. The illustrated model used one-month initial training data. The blue line indicates the mean prediction accuracy and the red line indicates the mean percentage of question skipping.

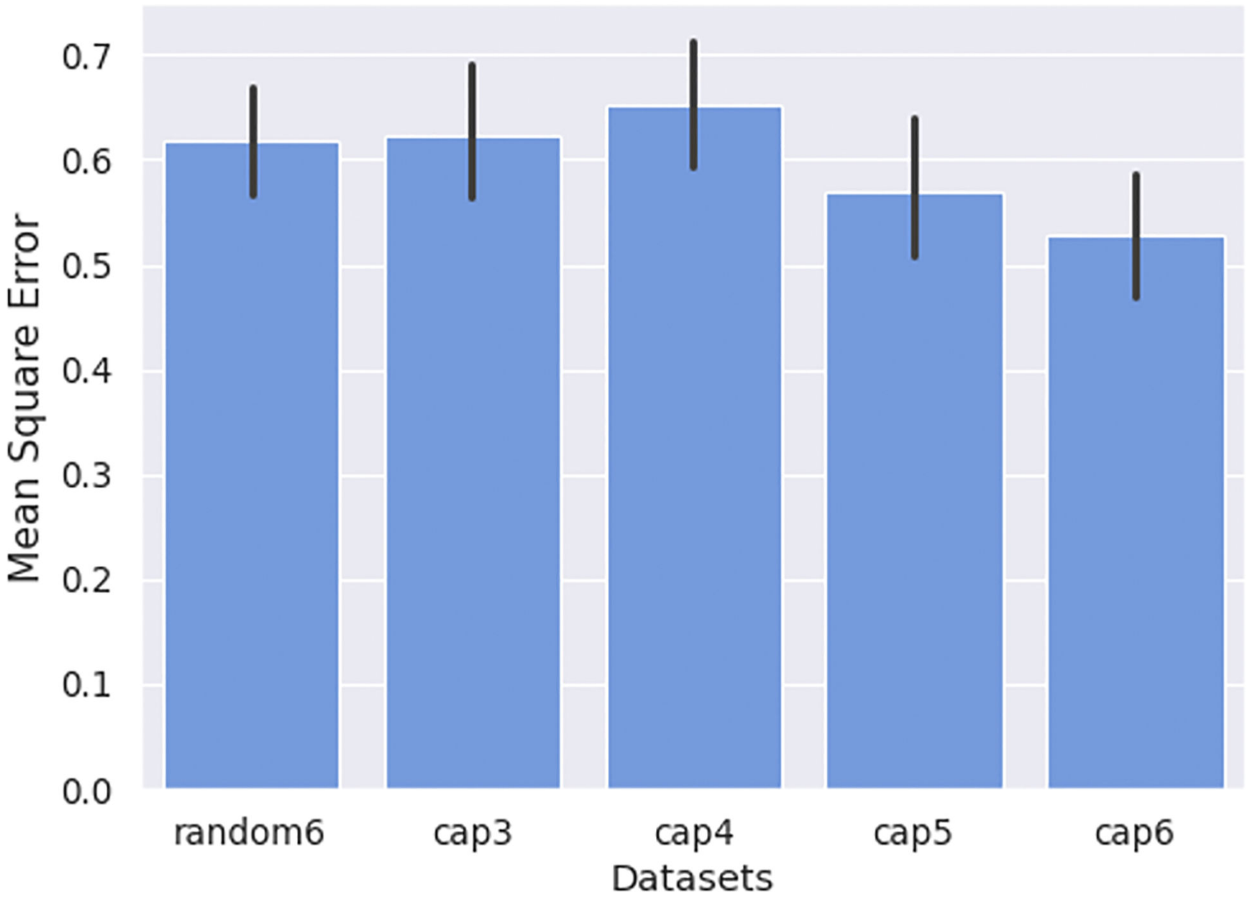


Figure 6.
Distribution of participants' average imputation errors in simulated Dataset 2.

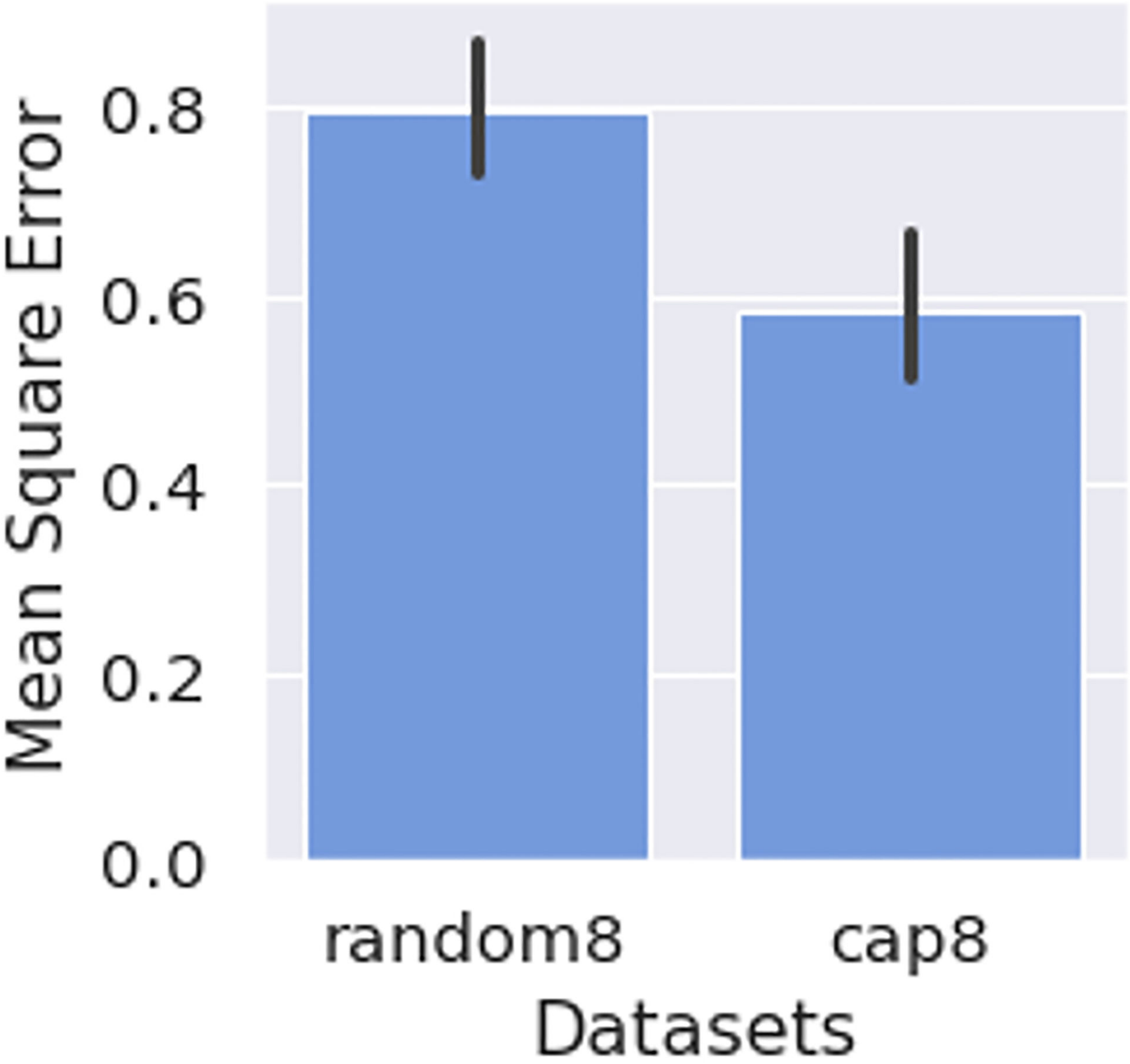


Figure 7.
Distribution of participants' average imputation errors in simulated Dataset 3.

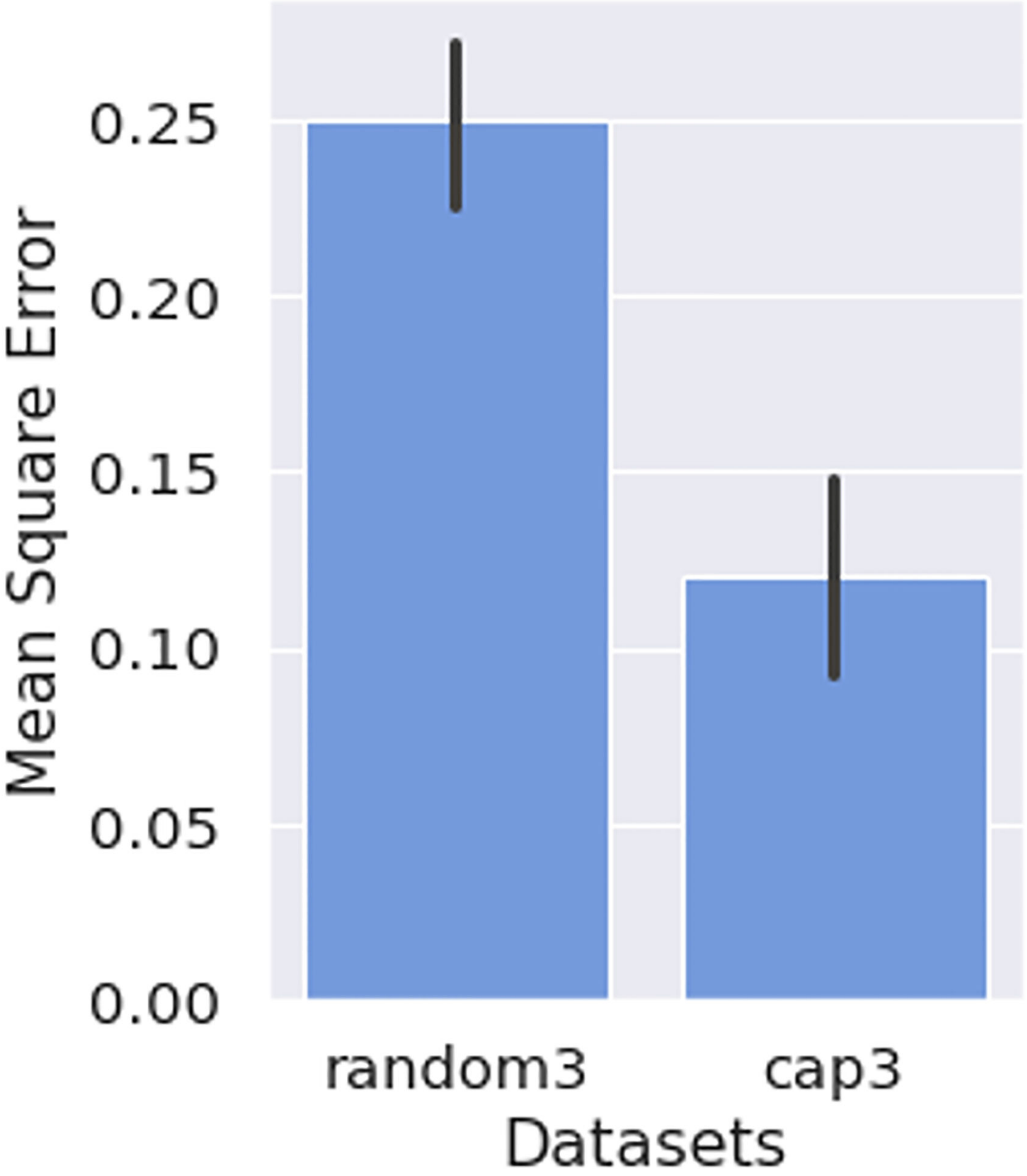


Figure 8.

Distribution of participants' average imputation errors in simulated Dataset 4.

Table 1.

Summary of real-world EMA survey datasets used in simulation studies

Dataset	Population	Duration	Prompt scheme	EMA question set
Dataset 1 (Daily EMA) & Dataset 2 (Hourly EMA) [46, 71]	Young adults (n=136)	One year	Prompted once every day before participants' anticipated sleep time / once every waking hour across four consecutive days every two weeks	12 questions on affect and feelings (daily summary/momentary)
Dataset 3 (Fixed-time EMA) [18]	College students (n=79)	14 days	Prompted four fixed times (noon, 3 p.m., 6 p.m., and 9 p.m.) each day	17 questions on mental health, social contact, and COVID-19-related behavior
Dataset 4 (Event-based EMA) [76]	Couples with or without personality disorder (n=228)	21 days	Active report immediately after every interpersonal interaction	31 questions from a hierarchical composite scale comprising 6 multi-item subscales on affect

Table 2.

Descriptive statistics about EMA survey datasets

	Dataset 1 (Daily EMA)	**Dataset 2 (Hourly EMA)**	**Dataset 3 (Fixed-time EMA)**	**Dataset 4 (Event-based EMA)**
Number of participants	120	134	74	228
Total number of responses	38,674	128,700	3,830	29,024
Mean number of responses per participant (SD)	322.3 (30.8)	960.4 (198.6)	51.8 (4.5)	127.3 (47.8)
Mean number of responses per participant per day (SD)	0.9 (0.1)	9.3 (1.8)	3.7 (0.3)	6.2 (2.1)

Table 3.

Experiments with lengths of initial full-survey training peri ods and number of response options.

		One month (30 days/8 burst days)		**Two months (60 days/16 burst days)**		**Three months (90 days/24 burst days)**	
		Five options	**Three options**	**Five options**	**Three options**	**Five options**	**Three options**
Daily EMA	Mean prediction accuracy % (SD)	63.7 (10.9)	80.1 (10.5)	69.4 (10.5)	83.1 (9.9)	71.8 (10.1)	**85.2** (8.8)
	Mean skipping percentage % (SD)	42.9 (14.7)	53.3 (15.8)	36.5 (16.8)	50.5 (16.1)	33.8 (16.6)	**48.1** (16.2)
Hourly EMA	Mean prediction accuracy % (SD)	73.2 (13.0)	87.1 (9.6)	76.5 (11.6)	85.7 (10.1)	77.5 (11.0)	**87.4** (8.6)
	Mean skipping percentage % (SD)	43.7 (18.2)	52.1 (20.4)	39.6 (18.0)	54.1 (17.8)	38.5 (18.3)	**51.9** (17.5)

Table 4.

Dependent t-test for paired samples: comparing participants' average imputation errors of random6 and cap3–6 (n = 134)

Dataset	Mean	SD	t-test (df=133)	Cohen's d
random6	0.6187	0.3053		
cap3	0.6229	0.3604	−0.1924	−0.01
cap4	0.6524	0.3812	−1.5921	−0.10
cap5	0.5704	0.3859	2.4992 *	0.14
cap6	0.5285	0.3377	5.3183 *	0.28

*Statistically significant at p<.05

Table 5.

One-week ahead prediction performance of simulated data collection for fixed-time EMA with five options.

	Variable-length stopping thresholds		
	0.2	**0.3**	**0.4**
Mean prediction accuracy of question skipped (%) (SD)	81.2 (13.5)	80.7 (12.1)	77.3 (11.5)
Mean percentage of question skipped (%) (SD)	45.0 (17.6)	56.3 (11.7)	71.7 (19.5)

Table 6.

One-week ahead prediction performance of simulated data collection for event-contingent EMA with five options.

	Variable-length stopping thresholds		
	0.1	**0.2**	**0.3**
Mean prediction accuracy of constructs skipped (%) (SD)	95.0 (7.1)	94.5 (6.5)	93.3 (6.3)
Mean percentage of constructs skipped (%) (SD)	48.2 (16.1)	60.3 (16.1)	68.3 (14.8)
Mean percentage of items skipped (%) (SD)	45.1 (12.4)	53.3 (11.7)	59.4 (11.4)