Gamification of Mobile Experience Sampling Improves Data Quality and Quantity

NIELS VAN BERKEL, The University of Melbourne
JORGE GONCALVES, The University of Melbourne
SIMO HOSIO, University of Oulu
VASSILIS KOSTAKOS, The University of Melbourne

The Experience Sampling Method is used to capture high-quality *in situ* data from study participants. This method has become popular in studies involving smartphones, where it is often adapted to motivate participation through the use of gamification techniques. However, no work to date has evaluated whether gamification actually affects the quality and quantity of data collected through Experience Sampling. Our study systematically investigates the effect of gamification on the quantity and quality of experience sampling responses on smartphones. In a field study, we combine event contingent and interval contingent triggers to ask participants to describe their location. Subsequently, participants rate the quality of these entries by playing a game with a purpose. Our results indicate that participants using the gamified version of our ESM software provided significantly higher quality responses, slightly increased their response rate, and provided significantly more data on their own accord. Our findings suggest that gamifying experience sampling can improve data collection and quality in mobile settings.

CCS Concepts: • **Human-centered computing → Field studies** • **Human-centered computing → Empirical studies in ubiquitous and mobile computing** • *Human-centered computing~Smartphones* • *Human-centered computing~Mobile devices*

General Terms: Location, Labeling, Crowdsensing, CSCW

Additional Key Words and Phrases: ESM, EMA, human behavior, sensing, experience sampling method, motivation

1 INTRODUCTION

The Experience Sampling Method (ESM) is used for human sensing and across many disciplines. Typically, the ESM aims to measure behaviour, thoughts, and feelings of participants *in situ*, thus providing detailed insights about both the participants' daily life as well as on the topic being studied as experienced by the participant

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), the European Commission (Grant, 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE).
Author's addresses: N. van Berkel: n.vanberkel@student.unimelb.edu.au; J. Goncalves: j.goncalves@unimelb.edu.au; S. Hosio: s.hosio@oulu.fi; V. Kostakos: v.kostakos@unimelb.edu.au.

2474-9567/2017/9-ART107 $15.00

DOI: http://doi.org/10.1145/3130972

[33]. This is achieved through repeated observations, allowing for phenomena to be assessed as they occur [44]. ESM study findings rely on both the quality and quantity of input, and thus motivating and incentivising participants is a crucial component of ESM studies [35]. The Ecological Momentary Assessment (EMA) is methodologically similar to the ESM, originating from the field of behavioural medicine, and the terms are now often used interchangeably [43]. In this paper, we refer to the term ESM.

While the ESM was originally intended to collect *personal* observations (*e.g.*, one of the first ESM studies analysed adolescent activity and experience [8]), it is increasingly adapted for other contexts. For instance, Connelly *et al.* [5] have used the ESM to explore knowledge hiding in organisations, and Van Berkel *et al.* [47] to quantify smartphone usage behaviour. More broadly, Burke *et al.* [1] suggested that user input in combination with sensor readings enables researchers to "*gather vital information about the built and natural environment that was previously unobservable*" [1]. At the same time, the use of sensor readings now allows researchers to "gamify" [11] the collection of data on smartphones, in hopes of increasing the quality and response rate of participants using smartphones. This gives rise to an important methodological question for researchers: should gamification be used in smartphone ESM studies, and if so, what is the effect on the collected data?

Determining the quality of human provided ESM data on smartphones is challenging. *Crowdsourcing* has been proposed as a way to rate the quality of human contributions, and is especially useful when the data concerns a publicly observable phenomenon. For instance, Hosio *et al.* [25] involve members of the crowd to approve the work of other crowd workers. Similarly, previous research has shown that crowdsourcing on public displays can be used to evaluate the quality of keywords that describe those locations through a majority voting scheme [16].

In our study, we developed two versions of the same ESM application: one with strong gamification elements [11], and one without. In a nutshell, our application asks participants to describe the environment around them. Because our study requires us to evaluate these ESM responses, we employ a mobile *game with a purpose* [49] to assess the quality of ESM entries. Therefore, our application allows participants to complete a task whereby they rate the contributions of other participants as either relevant or irrelevant. This enables us to calculate a 'score' for each contribution, measuring quality. To the best of our knowledge, our work is the first study to systematically investigate gamification as an incentive mechanism in mobile ESM studies.

2 RELATED WORK

2.1 ESM Response Quantity and Quality

Experience sampling is used to "*collect information about both the context and content of the daily life of individuals*" [22]. This is achieved by asking participants to answer a small set of questions at various moments over a period of time. By reconstructing the participant's context and provided answers, researchers can gain insight into the way these activities and contexts are experienced and perceived. Given the reliance of the ESM on participant responses, both the *quantity* and *quality* of these participant responses are important factors. These two factors ensure that the collected results are *ecologically valid* and capture "*the occurrence and distribution of stimulus variables in the natural or customary habitat of an individual*" [24].

Surprisingly, literature reports a wide range of participant response rates in ESM studies. Consolvo and Walker [7] asked their participants to complete 70 questionnaires on their mobile phone, resulting in an average of 56 completed questionnaires. Van Berkel *et al.* [47] presented a question at each phone unlock, resulting in an average response rate of 83.78%, ranging from 61.90% to 97.25% across participants. Yue *et al.* [50] investigated participants' willingness to voluntarily submit photographs when completing an ESM questionnaire: 30.80% of participants submitted at least one photo, with the total number of questionnaires completed with a photo attached was 3.50% and participatory levels being highest on the first day.

Yue *et al.* [50] counted the number of words and characters in the collected ESM responses as a measurement of the quality of the response, finding that questionnaires which had a photo attached correlated with higher data quality. Hicks *et al.* [23] show that 'participation quality' (defined as number of interactions) decreases as the battery level of the person's mobile device drops below a certain value. Furthermore, they categorise participants as 'power users' (using the device beyond study participation) or 'survey only users'. Participants using their personal device result in higher quality data when compared to those using study-specific hardware.

These studies indicate that the response rate significantly varies across contexts and means of presenting the questionnaire. In our study, we set to investigate if the rate can be improved through gamifying the ESM procedure. This is important because it would make such studies more efficient: more data is gathered in shorter time, leading into potentially more insights from participants.

2.2 Incentivising Participation

In experimental settings, incentives are a commonly used mechanism to entice participation, uphold retention rate, and ensure sufficient data quality [4,39]. Most ESM studies offer a form of monetary compensation to incentivise participants to submit more responses. Consolvo and Walker [7] offered $1 for each provided response, Yue *et al.* [50] offered participants $150 for 3 days of completed responses and $200 for 5 days of completed responses. More complex compensation structures are also applied, with Conner Christensen *et al.* [6] offering participants a flat fee for their efforts (approximately $20 for a week of participation), a physical remuneration for their efforts on a weekly basis (*e.g.* movie ticket), weekly drawings for smaller prices, and a final 'grand prize' at the end of the study. Lynn [35] shows that monetary incentives significantly improve participant compliance, with the strongest effect measured in diary completion (methodologically closely related to the ESM). Hosio *et al.* [25] introduced a market model, in which completed tasks are awarded a monetary compensation in accordance with the current market value. The manipulation of completed task rewards through price-setting proved efficient in steering participants to a specific task. Finally, a sufficient diversity in the available tasks ensured a sustained crowdsourcing market.

However, monetary incentives can also have inadvertently negative effects. For instance, an incentive of $250 for participation resulted in a low quality of collected data [45]. According to Stone *et al.* [45] this was the result of a participant base not intrinsically motivated to participate (solely participating for the financial reward). An alternative approach to motivating participants suggested by Larson and Csikszentmihalyi [33] is to form a "*viable research alliance*" between participant and researcher. Stone *et al.* [45] add that participants should be specifically encouraged to refrain from dropping out of a study, even if they forget to input data for one or two days. Khan *et al.* [30] combine the ESM with the Day Reconstruction Method in an attempt to improve quality of the collected responses and reduce data loss. Hsieh *et al.* [26] show that providing participants with a visualisation of their own data increases compliance rates. A different approach is described by Musthag *et al.* [39], concerning the use of micro-incentives to compensate participants throughout the study for completing a small task, as opposed to receiving a bulk-payment at the end of the study. Literature suggests three types of participant commitment in an online community [32]:

- *Affective commitment*: Following an emotional connection to the organisation or community.
- *Normative commitment:* A moral obligation to contribute, it is the righteous thing to do.
- *Continuance commitment:* Commitment originating from a certain incentive (either on individual or group level) that may be lost when leaving the community.

Lui *et al.* [34] classify community contributions into two factors: *individual* factors, consisting of both extrinsic motivation (*e.g.*, rewards) and intrinsic motivation (*e.g.*, reputation), and *interpersonal* factors (*e.g.*, liking or affiliation). Extending this, various incentive systems have been developed to encourage user

participation. Farzan *et al.* [13] consider the following incentive systems; incentivising with rewards, incentivising by explaining community benefit, incentivising by goal-setting, incentivising by reputation, and incentivising by providing self-benefit. Morschheuser *et al.* [38] investigate the effect of gamification on crowdsourcing and report positive effects (*e.g.*, increased engagement, quality).

Goncalves *et al.* [17] empirically validate a number of motivational approaches in a ubiquitous crowdsourcing setting. Their results show participant motivation can be influenced to elicit higher quality contributions and increase voluntary participation. Altruistic motivation alone is typically not sufficient of a motivator to convince people to contribute. Manipulations such as psychological empowerment [19], location cues [18], and increased enjoyment [29] all helped to increase participation - while simultaneously offering additional benefits (*e.g.*, increased quality of tasks, increased sense of urgency). Hence, literature suggests that contextual capabilities of ubiquitous devices can be used to increase participation and attitude.

2.3 Gamification and ESM

Deterding *et al.* define gamification as "*the use of game design elements in non-game contexts*" [11]. Gamification has been explored in the context of mobile applications. Fitz-Walter *et al.* [15] developed a gamified application for university students to be used during their orientation. Parts of the application relied on user input to complete certain quiz items. The researchers report some participants applying a 'trial and error' approach to finish these quiz items (*i.e.*, loss of data quality). Furthermore, one of the obtainable achievements asked students to 'check-in' to a number of events. The majority of participants stopped using this feature after the achievement had been unlocked.

Investigation of gamification aimed to support ESM or EMA collected data is scarce. Machnik *et al.* [36] present *Crowdpinion*, a tool that allows researchers to setup gamified ESM studies. In their approach, participants become a partner in the research study (as suggested in [33]) and can gradually unlock information about the ongoing study results as they submit data and even contribute questions to the study based on personal interest. This raises concerns on data integrity and quality, as highlighted by the authors. Gamification in *Crowdpinion* is limited to triggering the participants' curiosity, and does not include other (potentially) motivational elements.

Gamification as an element of motivation has been applied more extensively in applications aimed at medical intervention or health behaviour change [9]. While gamified intervention applications have been shown to work in the short run (*e.g.*, over a 6-week period [27]), other literature indicates that this effect declined in the long run (*e.g.*, after one year [12]). Additionally, the practise of gamification has also been criticised. For instance, a challenge of gamification is that the introduction of extrinsic motivations (*e.g.*, badges, points) may crowd out any prior intrinsic motivation (*e.g.*, altruistic effort, interest) [31]. Similarly, gamification may lead to a lack of understanding of the broader context in which the application is used [42]. A literature review from Hamari *et al.* [20] indicates that gamification does have a positive effect on motivational affordances and behavioural outcomes, but also notes it is challenging to reliably compare results between studies.

3 STUDY DESIGN

We investigate how gamifying an ESM study influences the quantity and quality of participant responses. To measure the quantity of responses, our ESM software asks participants to describe their current environment. To measure the quality of ESM responses, we created a *game with a purpose* that requires participants to rate the quality of words that describe their current location. Each ESM questionnaire therefore consists of two tasks; 1) submit a word that describes the participant's current surrounding, and 2) rate words submitted by others. In both cases, participants are asked to reflect on their personal observation of the current physical location. We therefore consider both the task of word submission and word rating to be part of an experience sampling

'survey'. Our participants were required to install our *GeoOulu* application on their personal Android phones. To assess the impact of gamification on the participants' ESM responses, we developed two versions of the application that only differ in their gamification elements.

Our experimental design is between subjects, with two experimental conditions: gamified software, and non-gamified software. The outcome variables include: ESM response rate, ESM response quality, and number of ESM responses.

3.1 GeoOulu

Our Android application *GeoOulu* consists of four screens, as shown in Fig. 1. There are numerous differences between the gamified and non-gamified versions of our game, as summarised in Table 1.

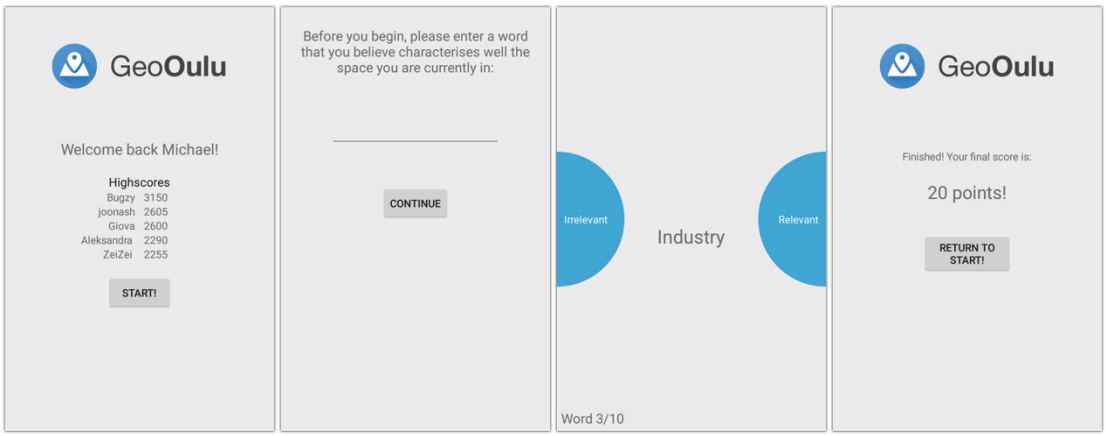

Fig. 1. (1) Application start screen, including leaderboard. (2) Submission of new word. (3) Rating of words. (4) Feedback (+ score) and option to return to the start screen.

- The start screen of the application welcomes the participant to the application and also allows the participant to choose a nickname upon initial launch of the app. The gamified version includes a scoreboard in this screen.
- The second screen asks participants to enter a keyword that describes their current location. The participant's location coordinates are recorded simultaneously (GPS + Wi-Fi localisation on Android).
- The third screen is the game's main screen. Here, participants are shown words previously submitted by other participants in the current location, and are asked to rate them as relevant or irrelevant. One game round consists of at most ten words that the participant needs to rate. The gamified version imposes time restrictions on completing the task.
- The final screen gives the option to return to the start screen and provides participants with feedback that the round was finished. In addition, the gamified version shows participants the score achieved in that round.

The purpose of the game is to rate each word as relevant or irrelevant to the current physical location of the participant. In the gamified version of the application, the words move from the bottom of the screen towards the top, and participants have 5 seconds to make a choice. If the participant does not rate the word before it has reached the top of the screen, it is automatically dismissed (and will not be shown to this participant again). In the non-gamified version, words appear centred on the screen without either animation or time restriction.

At the end of each game round, the data is submitted to the database and a score is calculated. Scores are based on the number of votes already cast on the selected word by other participants *at that point in time*. We apply a majority voting scheme [16], where agreement with the majority results in +10 points, disagreement with the majority results in -10 points, and in the case of an equal number of votes (including the 1st vote on a word) +5 points will be awarded. This majority voting scheme is scalable as it does not require a predetermined "ground truth". The accumulated score of each participant is calculated and stored in the database for both experimental conditions.

Table 1. Design of gamification elements between the two conditions as classified by game element level [11]

Element level	**Gamified**	**Non-gamified**
Game interface design patterns	Words to be ranked will animate from bottom to top of the screen.	Words to be ranked are fixed to the centre of the screen.
	A leaderboard is presented at the start of the game, showing the username and accumulated score of the five leading players.	No leaderboard is presented, and participants are unaware of their 'rank' to other participants of the game.
Game model	Following completion of a game round, the players is presented with the score achieved in that specific round.	Following completion of a game round, no score is presented to the participant. However, the metric is calculated and stored for further analysis.
Game design patterns and mechanics	A challenge of time is introduced, animating the words from bottom to top. Once the word reaches the top, the word will disappear and no vote is cast.	Words will remain stationary and will not disappear over time. Participant is free to take as much time as they want to rank each word.

3.2 Zones and Bootstrapping

Prior to the study we interviewed locals, who all had lived in Oulu for more than ten years, to determine a set of city landmarks as predetermined zones. The six zones are: the main shopping street, a railway station, a popular meetup park downtown, the main market square, the city library, and the university. The areas are widely known among locals, and we hypothesised they feature several distinctive characteristics. Besides the six predetermined zones, we defined a zone called 'Other', which covers all areas outside the predetermined zones.

To bootstrap the initial keywords used in the game, we recruited eight local volunteers (different from those involved in determining the zones) to submit keywords. Our volunteers had extensive knowledge of the area having lived in Oulu an average of 20.5 years, and thus being familiar with the locations. The volunteers submitted 466 words describing those zones. Then, the authors categorized the words and phrases using *emergent coding* [21], independently classifying all submissions into mutually exclusive themes, and then comparing notes until an agreement was reached on the keywords. The 10 most frequently occurring keywords from the volunteer submissions were used to bootstrap a word dictionary for each zone. In addition, we generated 10 random keywords per zone as a control condition, selected randomly from a language dictionary. Thus, each zone was bootstrapped with 20 words (10 from volunteers + 10 random). The 'Other' zone was bootstrapped with 20 random words.

3.3 Choice of Words

The main task of our game requires participants to rate the relevancy of words in regard to their current location. During our bootstrapping, the locations of our initial volunteer keywords were manually assigned to their respective zones. Our application uses an elaborate mechanism to choose the words that are shown to participants in any given game round:

- First, the system identifies words that were submitted in locations that are physically near the participant's current location. To overcome bootstrapping issues, words across the whole city were shown if necessary. For instance, unpopular locations may not have many words submitted nearby, so our system uses words that were submitted elsewhere in the city. Similarly, if a participant played multiple games in the same location, the system would eventually present words that were submitted at far away locations (over 500 m).
- Second, the system did not show to a participant a word that they had themselves submitted or rated previously at that location. This removed the potential bias of participants towards their own words. A word can, however, appear in a *different* location if someone else has submitted it there.
- Third, we carefully defined six specific 'zones' within the city. Within these zones, the distance of words did not matter and the whole zone is treated as a single conceptual location. This decision made it easy to avoid participants labelling the same word twice at the same location.
- Finally, for any game round we apply the rule that if five or less suitable words (*i.e.*, not submitted or previously ranked there by the participant) are found, the participant is unable to play the game. The user is presented with the following message; "*You have run out of playable words for this location. Please try a different location or come back later!*".

The criteria we define above achieve a number of goals. First, we ensure that all words eventually receive an approximately equal number of ratings, and avoid situations where a handful of words receive most of participant ratings. Second, our mechanism allows for spreading of words that are relevant to broad areas (*e.g.*, a beach covering a 2-km area). Our decision to show words from far away locations allowed for this possibility given the right keyword. Finally, we improve the quality and consistency of data since participants did not rate their own words, and participants did not play games that had too few words in them.

3.4 ESM Notification Scheduling

Notifications were used to prompt participants to complete the ESM questionnaire (*i.e.*, launch the app). A notification is sent to a participant when entering one of the six defined zones (*i.e.*, event contingent notification) using Android geofences, or when it has been two hours since the participant has last opened the application (*i.e.*, interval contingent notification). The Android geofence 'enter' event was chosen over 'dwell' to best match the ESM's philosophy of measuring events as they occur [44]. Using a 'dwell' timer to trigger notifications would result in the participant's most common locations being repeatedly triggered throughout the study (*e.g.*, home, work, university). Notifications are shown in the default Android notification menu, and interval contingent notifications are only generated between 08:00 and 22:00. The outcome of a notification may be:

- The participant accepts the notification.
- The participant explicitly dismisses the notification.
- The notification expires automatically within 15 minutes, or if the participant has left the zone.

3.5 Recruitment and Experimental Procedure

We recruited 24 participants using mailing lists of our university (16 males, 8 females; ages: 21-38 years old, mean = 27.00). Participants were required to have lived in Oulu for at least six months and to own an Android-based smartphone. On average our participants had lived in Oulu for 5.58 years (SD = 7.88), suggesting that they should have already amassed knowledge of the town and the zones we defined. In addition, our participants had a diverse range of educational backgrounds (*e.g.*, Accounting, Education, Plant Ecology, Wireless Communication).

We employed a between-subjects design in which half of the participants were assigned to the gamified condition and the other half to the non-gamified condition. Assigned conditions did not change during the study. In addition, we asked participants to complete a pre-study questionnaire, focusing on demographic information and mobile phone usage. One of the questions asked participants whether they use their mobile phone to play games (with the possible answers of 'never', 'sometimes', and 'regularly'). We controlled for this variable by balancing participants across the two experimental conditions based on their response. This resulted in 2 participants who 'regularly' play games on their mobile phone (1 in the gamified condition), 15 participants who 'sometimes' play games on their mobile phone (7 in the gamified condition), and 7 participants who reported to 'never' play games on their mobile phone (4 in the gamified condition). Participants in the gamified condition had an average age of 26.50 (SD = 2.32), compared to an average age of 27.50 (SD = 4.64) in the non-gamified condition. A Welch's unequal variances *t*-test indicated no significant difference between conditions ($t = 0.67$, $df = 16.16$, $p = 0.51$). As for their duration of stay in Oulu, participants in the gamified condition reported an average of 4.63 (SD = 6.41) years compared to 6.54 (SD = 9.31) years for participants in the non-gamified condition. A Welch's unequal variances *t*-test indicated no significant difference between conditions ($t = 0.59$, $df = 19.52$, $p = 0.56$).

To minimise learning effects, we held individual intake training sessions. During these sessions, we installed the application and guided participants through all screens of the application (Fig. 1). While doing so, we provided several examples of both 'good' and 'bad' keywords for the location of the intake session (university). For example, the keyword 'knowledge' was provided as a good example of a word that described the university location. The keyword 'couch' was given as an example of a bad keyword, since it is restricted only to that specific room. We encouraged participants to take a broad view when submitting and ranking words, considering not only their current isovist view (*e.g.*, the current room) but also the setting they are in. In order to keep the explanation offered to participants as similar as possible between the two conditions, we did not discuss the implementation of the scoring mechanism (only visible in the gamified application). Participants in the non-gamified condition were therefore never aware of any score being calculated, as this information was never presented to them in the application. Although participants in the gamified condition were aware that a score was being kept, we did not explain how this score was calculated and whether it was based on their own submission or on their rating of words submitted by others. During the data collection phase, participants could freely launch and use our application at any time, and in addition received notifications according to our scheduling criteria.

After the data collection phase ended, we invited participants for a one-on-one debriefing session with a researcher. During this debriefing, participants were asked to complete an exit questionnaire, which contained, *inter alia*, questions regarding notifications, gamification elements, and potential problems the participants might have experienced. As a compensation for their participation, each participant received two cinema vouchers.

4 RESULTS

Data collection was limited to a three-week period, as the quality of ESM responses is known to deteriorate after a period of 2-4 weeks [45]. During the study, 811 game rounds were completed, a total of 1197 words were submitted, along with 7902 word ratings. The number of game rounds completed is limited by the availability of new words for people to play – explaining the difference between the number of word submissions and total rounds of rating words. Most rounds were completed in the 'Other' zone (623), followed by the market square (41), downtown (41), university (40). The least popular location was the railway station (13). Fig. 2 shows a) the total number of completed game rounds, b) participants' average score per round, and c) the total number of submitted words, broken down by study condition and study day.

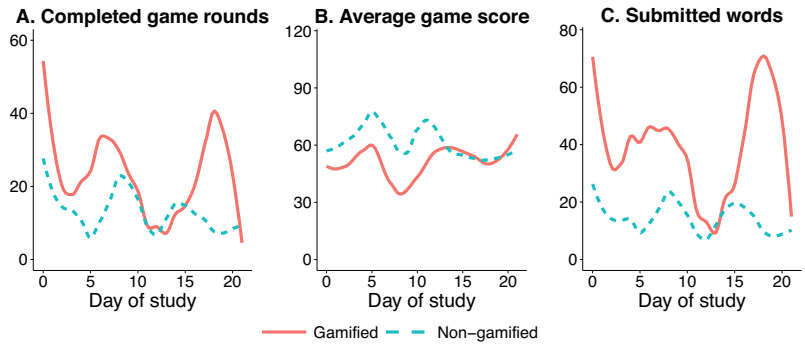

Fig. 2. A: Number of completed game rounds. B: Average game score. C: Number of submitted words. All values shown over the duration of the study per condition.

4.1 Quantity of Responses

Participants in the gamified condition completed 516 game rounds and submitted 877 words, versus 291 rounds and 320 words in the non-gamified condition respectively. The total number of game rounds completed per week did not change considerably over the duration of the study, with 298, 238, and 271 number of game rounds completed for week one, two, and three respectively. No statistically significant difference was found between the participants' tendency to play mobile games on their phone and their final participant score, as determined by a one-way ANOVA ($F(2,21) = 0.06$, $p = 0.94$). We provide a boxplot of a) the total number of completed game rounds, b) participant's average score per round, and c) the total number of submitted words in Fig. 3 to visualise the distribution between conditions.

Of the 1197 words submitted, 610 words were unique submissions. A Kruskal-Wallis test on the number of words submitted (per day) per condition shows a statistically significant difference between the gamified and non-gamified condition ($H(1) = 12.98$, $p < 0.01$), with a respective submission mean of 5.38 and 2.89 per day. The majority of submitted words was unique across the study, *i.e.* they were inserted only once (N=490). However, 120 words were inserted more than once, implying participant agreement on certain characteristics of the zone. Fig. 4 depicts the breakdown of *reoccurrences*, *i.e.* how many times a word was inserted again by other participants after its initial entry to the dictionary.

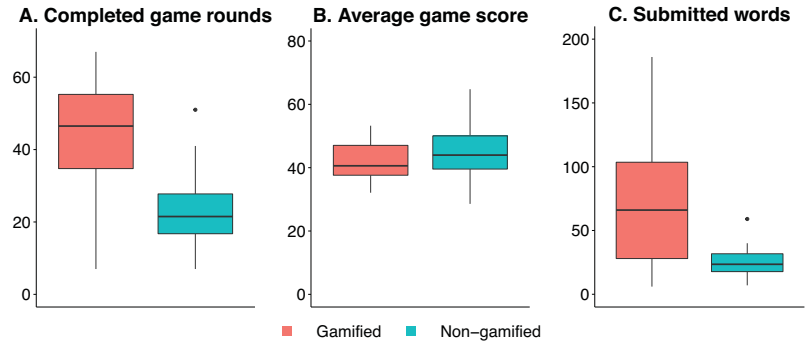

Fig. 3. Boxplot showing distribution between gamified and non-gamified participants.

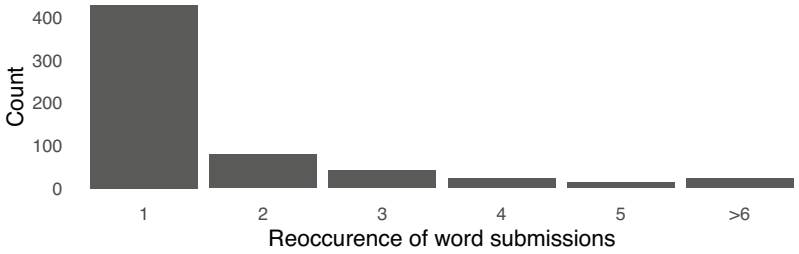

Fig. 4. Number of reappearing word submissions.

4.2 Quantity and Location

Fig. 5 shows the number of game rounds completed in each of the six zones in our study, split per condition. Participants in the gamified condition completed more game rounds (126 vs. 58 when considering the aggregate sum). A two-factor ANOVA shows a statistically significant interaction effect between the experimental condition and the zone on the number of completed game rounds ($F(1, 176) = 5.55$, $p < 0.01$).

The number of game rounds completed per day averaged 5.38 and 3.03 for the gamified and non-gamified conditions respectively. The effect of study condition on number of game rounds completed was statistically significant ($F(7, 176) = 13.34$, $p < 0.01$), as was the effect of location ($F(7, 176) = 89.71$, $p < 0.01$). A post-hoc Tukey HSD on the different locations shows a statistically significant difference between the 'Other' zone and all predefined zones ($p < 0.01$) for all combinations, with the 'Other' zone having a higher number of completed game rounds compared to all other zones. There was no significant difference between any combination of predefined zones. Finally, as shown in the map presented in Fig. 6, the gamified condition led to an increased density of contributions, though no considerable difference in geographic coverage can be detected between conditions.

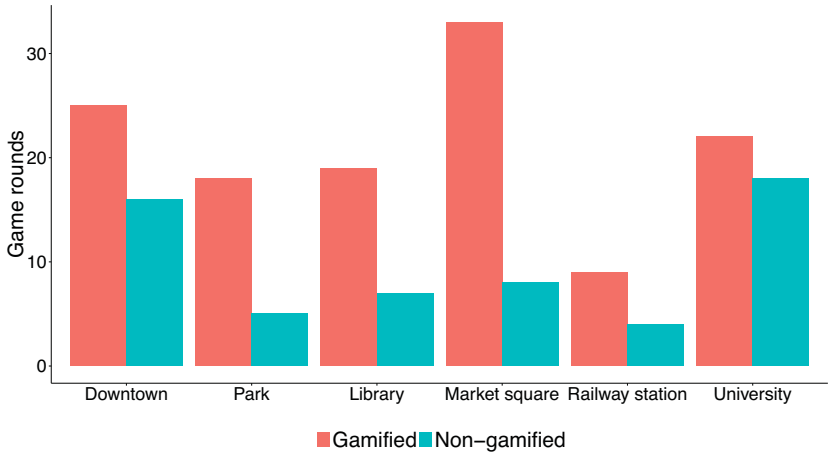

Fig. 5. Completed game rounds per predefined zone.

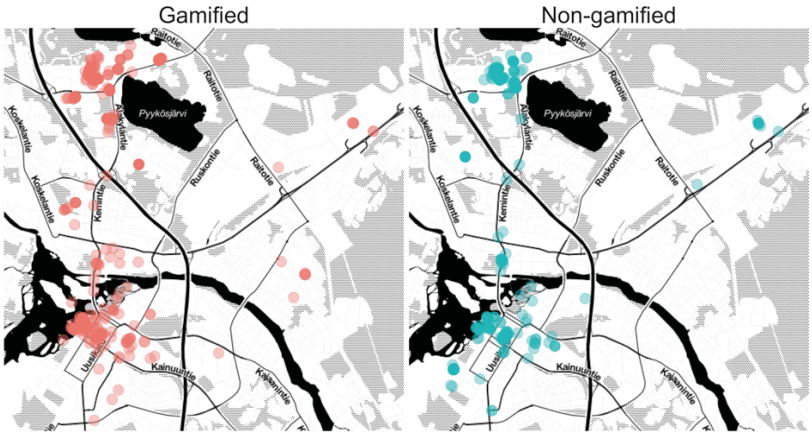

Fig. 6. Location of completed submissions

4.3 Application Entry and Notifications

We compare the various ways in which participants could launch the game and contribute:

- Clicking on the application icon to launch it.
- Accept a notification triggered by 2-hour interval rule.
- Accept a notification triggered by entering a zone.
- 'Replay game' after completing a game round.

Table 2 provides a distribution of these frequencies. A chi-square test shows a statistically significant difference between entry and study condition ($\chi^2 = 11.98$, df = 2, p-value < 0.01). For this test, we do not consider the 'Replay game' entry as it always follows one of the other categories. Note that for both conditions it is possible that incoming notifications primed participants to open the application directly from the homescreen rather than by opening a notification. In our data, this would register as launching the application through the application icon.

Table 2. Distribution of application entries, *leading to completion of a game round*

Entry	%	Gamified	Non-gamified
Application icon	42.87%	235	111
Notification (time)	25.65%	111	96
Notification (loc.)	10.53%	49	36
Replay game	20.94%	121	48
		(0.31 replay ratio)	(0.20 replay ratio)

Table 3 shows a breakdown of all 1908 notifications sent during the study, grouped by condition. The difference in the number of notifications between the study conditions is the result of multiple factors including participants' mobility, daily routines, and motivation to travel. We note that these notifications did not all lead to completed game rounds (*e.g.*, ignoring notification), explaining the lower total number of notifications reported in Table 2. A chi-square test shows a statistically significant difference between study condition and notification interaction ($\chi^2 = 44.96$, df = 3, $p < 0.01$). Pearson residuals show an overrepresentation of dismissing notifications and expiring (loc.) notifications for the non-gamified condition (respectively 3.94 and 2.53).

Table 3. Overview of notification interaction

ESM Notif.	Gamified	Non-gamified	Total
Answered	302 (28.33%)	204 (24.23%)	506 (26.52%)
Dismissed	71 (6.66%)	121 (14.37%)	192 (10.06%)
Expired (time)	680 (63.79%)	487 (57.84%)	1167 (61.16%)
Expired (loc.)	13 (1.22%)	30 (3.56%)	43 (2.25%)
Total	1066 (100%)	842 (100%)	1908 (100%)

Participants responded to a total of 506 notifications, giving a response rate of 26.52%. Overall, 10.06% of total notifications were actively dismissed by participants. The majority of notifications (61.16%) was dismissed due to expiration of the notification timeout (15 minutes), and 2.25% of notifications were dismissed due to the participant leaving one of the predetermined areas.

4.4 Quality of Responses

While most game rounds were completed in the 'Other' zone, we exclude it from the analysis of *quality* of the submissions. This is because the study, by design, yields heterogeneous words in this zone, as participants were free to roam anywhere. For the same reason words submitted in this zone ranked considerably differently from words submitted in the six pre-defined zones (mean score of 42.34 for 'Other', whereas all other zones have a mean score ~ 80, see Fig. 9). In addition, our volunteer submissions did not cover the 'Other' zone due to the fact that this area was not limited to a specific location. It was therefore not possible to provide a set of location-specific descriptive words. This lack of volunteer words potentially skews the results, as participants did not have 'example' words.

To measure the quality of submitted words, we rely on the ratings obtained through our game (crowd-rating). Participants classified words to be either relevant or irrelevant, with a participant unable to classify a word they have submitted themselves. Each word's score is calculated by dividing the number of relevant votes by the total number of votes for that word. We distinguish between words submitted by our volunteers (bootstrap),

the random words (bootstrap), words submitted by participants in the 'gamified' condition, and words submitted by participants in the 'non-gamified' condition.

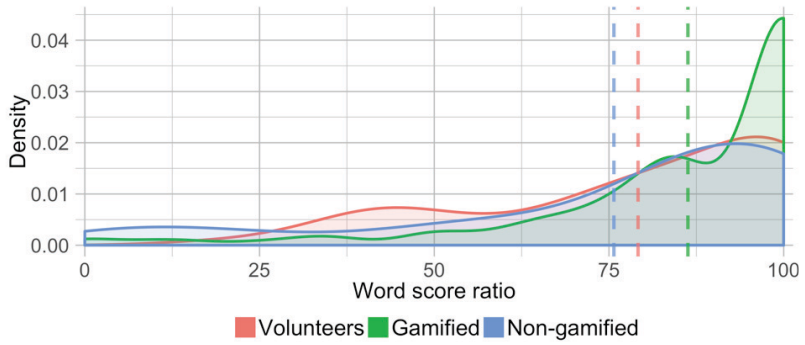

Fig. 7. Density plot of word score ratio of submitted words.
Average score per *human word source* indicated by vertical lines.

As expected, the 'random' words have the lowest average score ratio (3.49). Fig. 7 shows a density plot of the three different *human word sources* and their respective relevance ratio (volunteers, 'gamified participants', 'non-gamified participants'). A Kruskal-Wallis test revealed a statistically significant difference between the word sources (even when excluding the 'random' source), ($H(2) = 8.05$, $p = 0.02$), with the mean acceptance ratios shown in Table 4. A post-hoc test using Mann-Whitney tests with Bonferroni correction showed a significant difference between gamified and non-gamified word source ($p < 0.05$). The boxplot in Fig. 8 shows the distribution of participants' mean word scores between conditions.

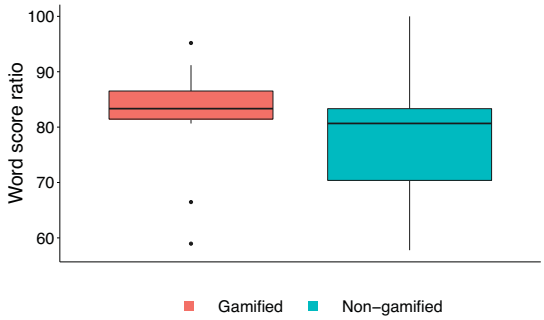

Fig. 8. Boxplot showing distribution between gamified and non-gamified participants.

Table 4. Word score ratio per submitted word source

	Mean score ratio	**St. Dev.**
Random dictionary words	3.49	9.02
Volunteers	79.16	22.58
Gamified	85.31	21.62
Non-gamified	76.28	28.52

4.5 Quality and Location

We also compare the quality of the submitted words (the crowdsourced relevance ratio) per zone, as shown in Fig. 9. A Kruskal-Wallis test shows a statistically significant difference in relevance ratio between the different zones, $\chi^2(6) = 384.39$, $p < 0.01$. A pairwise comparison with Bonferroni correction shows a statistically significant difference between all pre-assigned zones and the 'Other' zone – no statistically significant difference was found between the remaining zones.

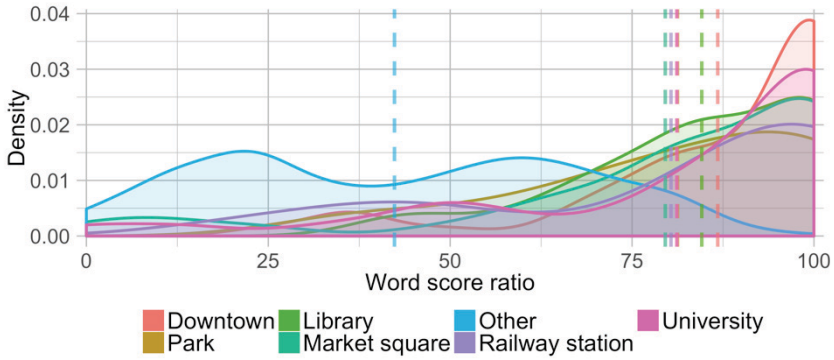

Fig. 9. Density plot of word score ratio of submitted words.
Average score per location indicated by vertical lines.

4.6 Participant Game Scores

Average scores per completed game for the gamified and non-gamified conditions were 53.10 and 60.26 respectively. A Kruskal-Wallis test on the achieved score (per day) per condition shows no statistically significant difference between the two conditions ($\chi^2(1) = 1.14$, $p = 0.29$). Average total score at the end of the study for gamified and non-gamified participants was 557.50 (SD = 409.84) and 291.25 (SD = 329.10) respectively. This discrepancy between the slightly lower average score per game round and the much higher total score for participants in the gamified condition can be explained by the higher number of completed game rounds (see Table 2). Lowest and highest participant scores were 0 and 1460 respectively. Of the eight best performing participants, six were in the gamified condition. Of the eight participants with the lowest score, two were in the gamified condition.

4.7 Application Entry and Quality

Table 5 shows the quality of submissions (word quality score ratio and participant score) per entry type. The 'Replay game' entry can follow any of the other entry types. Quality of word submissions remains roughly equal across entry types. The mean scores per completed ranking show a considerable lower score for entries that originate from a time notification. Low number of time notifications is due to the study design, in which a *location* notification was sent when entering one of the predefined zones. Table 5 shows that notifications are responsible for enticing most interactions – this as opposed to Table 2 (which includes the 'Other' zone), where the application icon resulted in most interactions.

Table 5. Word quality and game round score following different application entry categories

	Word quality (mean)	**Game score (mean)**	***n* gamified**	***n* non-gamified**
Application icon	90.55	50.70	33	24
Notification (time)	93.23	32.78	7	2
Notification (location)	89.05	54.19	41	27
Replay game	89.41	66.30	45	5

5 DISCUSSION

ESM studies rely on human participants providing high-quality responses. While the importance of both the quantity and quality of these responses has been investigated previously [33], surprisingly little work examines motivators other than financial stimuli (studied for example in [7,25]). One could even argue research has regarded participating in ESM studies as an annoyance that can be potentially reduced through intelligent notification techniques [14]. Our results indicate that gamification of the ESM increased participation motivation and led to a considerable higher number of pro-active data contributions. We now discuss both the quantity and quality of contributions and the gamification elements we used.

5.1 Gamification Improves Quantity of Contributions

Participants in the gamified condition submitted significantly more words (+174%) and completed a significantly higher amount of game rounds (+77%) when compared to participants in the non-gamified condition. When considering the number of completed game rounds over the predefined zones (Fig. 5), we found that participants did not only complete more rounds in all zones, but were more willing to travel to zones not part of their daily routine (*i.e.*, the park, library, and railway station) – this observation was also reflected in participant interviews.

We note that the difference between conditions could be even larger, but that the restriction on only rating words once per location has deterred participants from starting the application altogether and thus not submit any new words. One of the participants notes "*if the app didn't restrict me [on the] number of plays for the location, I would keep playing more*" (P21). This is supported in the collected data, whereby non-gamified participants submitted only 29 more words than completed game rounds (additional 10%), versus 361 more words than completed game rounds (additional 70%) for participants in the gamified condition. All these submitted words could have resulted in completed game rounds, were it not for the aforementioned restriction.

The stark difference in number of completed rounds in the 'Other' zone versus all predefined zones is the result of a positive feedback loop. In the 'Other' zone, a sufficient number of words was often available, resulting in a continuous flow of new contributions. The predefined zones were unable to create or sustain such a critical mass, leading to exhaustion of the available words. This effect can be observed as an (unintended) extra gamification element, in which the number of available games (and thus available points) were limited – requiring more active effort from participants to be able to complete game rounds.

As shown in Table 5, the gamified condition did not lead to considerable higher response rates on incoming notifications. Participants in the gamified condition were, however, more willing to both proactively launch the application themselves, and to provide multiple rounds of input in succession. This latter finding supports the findings by Jung *et al.* [28], who show that providing feedback results in a higher number of contributions – especially in the case of contributions using pseudonyms (as opposed to anonymised contributions). In addition,

participants from the non-gamified condition were more likely to actively dismiss incoming application notifications, indicating a lower level of interest or commitment to the application.

Following the terminology from Chang *et al.* [3], we consider a participant proactively opening the application to be engaged in participatory sensing – whereas reacting to notifications is considered context-triggered *in situ* sensing. Literature suggests that the number of participatory sensing contributions can be increased through gamification [46]. These results are supported by our findings.

With an average response rate of 26.52% to notifications, the study's response rate appears relatively low compared to other studies (*e.g.*, [5] with 65%, and [41] with 52%). The current study contained a total of 1908 notifications, an average of 79.5 notifications per participant over the duration of the study, or an average of 3.79 notifications per day. The aforementioned studies have very different study designs (1 daily survey over five successive working days [5], 10 tasks per day over a four-week period [41]), making it difficult to compare and discuss response rates between studies. Multiple factors explain the relatively low average response rate. The application automatically dismissed notifications after a timeout of 15 minutes, as well as when leaving the premise of one of the predetermined locations. Therefore, a total of 63.41% of notifications were automatically discarded by the system, while participants actively dismissed only 10.06% of notifications. Therefore, the apparently low response rate can be explained by our strict expiry mechanisms (time and location) while participants only dismissed a relatively low percentage of ESM notifications.

Furthermore, the number of tasks required to be completed was relatively large in comparison to other ESM studies. The acceptability of a certain number of notifications is highly dependent on the time and effort required to complete each questionnaire [7]. In comparison to a previous study [47] for example, where participants were immediately presented with a single question on screen unlock, our study required considerable more effort from participants. Lastly, the nature of the application could have deterred participants from replying to some notifications; "*[I] didn't feel the need to answer about the same place as I have done just before.*" (P07). Therefore, we conclude that the design of the ESM's contingency, in which a participant's change in location would result in a new notification (event contingent), has led to an overflow of irrelevant notifications. Future studies should embed more intelligence in their event driven notifications, for example by inferring whether a participant recently answered a question related to the triggered event.

The number of completed game rounds are not equally distributed over the duration of the study (Fig. 2 - A). Given the fact that this effect is observable in both the gamified and non-gamified condition we believe this is the result of the game design (participants are only allowed to rate each word once) rather than participants increased interest when their position on the leaderboard is endangered. Other external factors such as the weather could also have influenced the participant's willingness to contribute.

5.2 Gamification Improves Quality of Contributions

The quality of the words submitted by gamified participants was significantly higher than those of the non-gamified participants. Word quality of gamified participants was higher than that of our bootstrapping volunteers, albeit not statistically significant, this difference is remarkable as our volunteers are local experts. A possible explanation for this might be the that our bootstrapping volunteers did not provide their contributions *in situ*, although this needs further exploration. Contributions made by participants in the non-gamified condition had on average a slightly lower score than words given by our volunteers. Since we did not explain how scores were calculated, those in the gamified condition might have been under the impression that the quality of their submission influenced their score – increasing their motivation to provide high quality submissions.

Despite participants in the gamified condition submitting words of higher quality, they were unable to achieve a higher average game score. In fact, their average game score per submission round was lower than

those in the non-gamified condition (although not statistically significantly). The gamification element of time restriction, in which participants in the gamified condition were given a maximum of 5 seconds per word to decide on the classification of a word, might have influenced this result. Although the log data shows that a participant's word timed out during gameplay just once in one of the predetermined zones, the added pressure of the time constraint might have influenced the participants' judgement. Cechanowicz *et al.* [2] include a timer as gamification element in an online questionnaire, and note how it might have led to shorter answers on free form text input. They therefore consider time-based game elements a fundamental trade-off between data quality and the effect of gamification. It can also be considered that participants in the gamified condition were aiming for a high final score, but did not so much aim for a high score in individual game rounds; "*I tried to use up all the words the app provided from the places I visited, to gain maximum points.*" (P09).

We highlight the difference in scores obtained by our participants following their entry into the application. Although only including a limited number of samples, Table 5 shows a lower mean score for those participants who opened the application following a time notification. This effect is not present for those opening the application through a location notification. Mehrotra *et al.* [37] suggest that ill-timed notifications might result in dismissed or hastily completed questionnaires, reducing quality of collected user data. A sensor based technique is proposed that predicts user's receptivity to answer incoming ESM notifications. Although our data does not allow for conclusive findings on this matter, it provides further indication of the importance of interruptibility-related work on manual data collection (*e.g.*, Visuri *et al.* [48]).

Lastly, we compare the quality of submitted words with the study presented in [16], which features a similar majority voting scheme and contained various gamification elements. Converting the obtained overall participant score from that study to our ratio results in an average word ratio of 66. This word ratio is lower than that obtained by both our non-gamified (76.28) and gamified (85.31) participants. The system used by Goncalves *et al.* [16] to collect these words ran on a public display, and was unable to keep a personal record of word submissions. Since contributions were not linked to an individual, participants might have felt less enticed to submit quality contributions.

5.3 Perceptions of Gamification Elements

Analysing the exit questionnaire, we identify the leaderboard as a crucial element in fostering competition and encouraging participants to submit data. "*I'm more competitive because of the leaderboard.*" (P19), and "*I tried to use up all the words the app provided from the places I visited, to gain maximum points*" (P09) indicate how participants in the gamified condition experienced the leaderboard. At the same time, some of the participants indicated that the leaderboard functionality did not influence their application usage. This is in line with earlier findings on gamification [10]. Standard deviation of the number of game rounds completed daily is considerably higher for the gamified condition (3.45 vs 2.47 respectively). This might indicate that gamification motivates a set of the participants, while it might have no direct effect (or even act as a deterrent) for others.

Participants' motivation, or rather the lack of it, has been mentioned as one of the challenges for the ESM [43], influencing whether someone will successfully complete an ESM study. Our data shows that the use of gamification elements had a positive effect on participants' number of contributions, and the exit questionnaire results confirm the effect on motivation; "*[...] when someone was before me [in the leaderboard], I was more motivated to gain points and beat the other players*" (P21). Although gamification proved beneficial in increasing the number of submissions in this case study, the technique is not suitable for all studies. This puts gamification of the ESM in line with other methods focused on increasing participant contribution suggested in the literature. Some examples are: reducing the number of questions [43], engaging participants to feel part of the study [33], complex remuneration structures [6], offering feedback visualisations [26], or simply extending the duration of

the study [6]. It is up to the researcher to determine which method is suitable to be applied in the concerning study.

When asked what could motivate participants to submit more data, responses were diverse. Two of our non-gamified participants suggested gamification elements without primer: "*point system in an interconnected network of users*" (P13) and "*there could be some kind of 'reward' or points based system*" (P18). Others believed the timing of questionnaires to be of importance: "*the phone can give notification to remind, when I am trying to relax*" (P05).

While the application showed a feedback screen following the completion of a round to all participants (Fig. 1 - 4), those in the gamified condition received more direct feedback on their performance in the form of an obtained score – something naturally not included in the non-gamified condition. The inclusion of additional feedback could have resulted in an increased participant engagement, as for example observed in [40] where an automated voicemail system used to deliver responses occasionally changed its utterances. Since participants never rated the same word in the same zone and were not made aware of which words were labelled correct or incorrect, we believe that the display of the score did not result in any learning effect to these participants.

While we were unable to directly compare the effect of the four different gamification elements (Table 1), participant answers from the post-study questionnaire indicate that gamification (leaderboard) and, to a lesser degree, individual score, were considered to be main influencers: "*I was somehow motivated in being in the top position. That's why I've started many times the application, even more than once in two hours.*" (P19). The challenge introduced to our gamified participants through the use of a timer has likely introduced unintended pressure, resulting in a lower average score per round. Although this cannot be concluded with certainty, other studies indicate a similar risk in time-pressure operations [2].

5.4 Motivation in Experience Sampling

As suggested by the related work on experience sampling, monetary incentives are the most commonly used motivators to both attract study participants and entice continuous data submissions (*e.g.*, minimum number of data submissions to obtain a study reward, micro-payments per submission). While Lynn [35] shows that these monetary incentives do have a positive effect on participant compliance, there are also reports on the negative effects of momentary incentives such as a low quality of submitted data [45]. While other motivational methods have been discussed in the literature (*e.g.*, constructing a "research alliance" [33], checking on the participants' contributions from time to time), the effect of these methods have not been empirically studied. The work of Hsieh *et al.* [26] on visualisation as a method to improve response rate is an exception to this. The study design of that work is similar to our own, where a control group is introduced and provided with the exact same notification schedule and questionnaire configuration – only removing the tested motivational stimuli.

Compared to the current literature in ESM motivation, our work provides a new way for researchers to motivate their study participants. While the concept of gamification in itself has been widely tested in other methodologies, the nature of experience sampling – in which both quantity and quality of user submissions are challenging factors – does require for gamification to be applied in a coherent manner. As our results show, 'gamified' participants outperformed those in the non-gamified condition on both submission quantity *and* quality, but achieved a lower average score per game round. We believe this is the inadvertent result of a game mechanism, in which participants experienced (time-)pressure to answer the presented questions. Researchers should take note of this in their own study designs, avoiding game mechanics that can adversely affect submission quality (*e.g.*, time pressure, distracting visuals / noises, and sensory information overload).

5.5 Limitations

First, while our study allows for obtaining insight on the quality of user contributions, we acknowledge that such assessment is not always feasible for ESM studies that aim for collecting individual accounts, and not for aggregating opinions like ours. Thus, our method for determining response quality does not generalise to all ESM studies. Second, our participant population consisted mostly of tech-savvy and relatively young individuals who all had experience in using smartphones and responding to phone notifications. We have yet to investigate how gamification would in practice affect if the respondents were for example elderly people, or if the study was run in entirely different cultural or geographical contexts. However, our methodology can be directly replicated for those contexts. In addition, we did not explicitly indicate which ratings were 'correct' or 'incorrect' as this information was irrelevant for the purpose of our data collection. However, including such feedback could affect ESM responses. Lastly, our study is limited to a set of four gamification elements (Table 1), as such, other gamification elements were not explored. We for example did not include any (gameplay) sounds as it is common for smartphone users to mute their device during daily usage.

6 CONCLUSION

In this paper, we systematically investigate the effect of gamification on ESM responses through a user-study with 24 participants. We use a location-based *game with a purpose* to collect and rate submissions. This game with a purpose asked participants to submit and subsequently rate labels describing their current location. Through a majority voting scheme, this type of collected 'location dictionary' can prove useful for a variety of purposes including city planning, commercial activities by local shop owners, and localised search engines. Our results indicate that gamification enticed participants to provide 174% more word submissions and rate 77% more submissions of other participants. Given the nature of the ESM, this is an important finding since a higher number of submissions allows for a more complete overview of the participant's daily experiences, and behaviour. This increase in contributions was primarily the result of user initiated activity. Finally, the submissions from participants in the gamified condition received significantly higher ratings than those in the non-gamified condition. Given the limited sample of 24 participants, our results offer the starting point for a plausible discussion on this matter within the community interested in this subject.

The use of a leaderboard and score as gamification elements proved to be the most effective in motivating participants. Our use of a 'time challenge' as a gamification element however may have had a negative effect, as those participants in the gamified condition had a lower average score per completed game round. Although gamification is not suitable for all ESM studies, our results show that it does entice participants to provide more information throughout their day-to-day activities. A combination with other sources of participant motivation (*e.g.*, visualising contributions [26]) could be considered useful. The downsides associated with motivating ESM participation (solely) through monetary incentives [45] call for the investigation of other potential incentives. This paper shows the potential for gamification in an experience sampling study, describing both its effect on response quantity and quality. Given that these motivators have been shown to work, researchers can apply this incentive in their own studies, or extend and improve the work by investigating other potential incentives.

7 REFERENCES

[1] Jeffrey A. Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy and Mani B. Srivastava. 2006. Participatory sensing. In *First Workshop on World-Sensor-Web: Mobile Device Centric Sensory Networks and Applications at Sensys '06.*

[2] Jared Cechanowicz, Carl Gutwin, Briana Brownell and Larry Goodfellow. 2013. Effects of Gamification on Participation and Data Quality in a Real-world Market Research Domain. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, ACM, 58-65. DOI: https://doi.org/10.1145/2583008.2583016

[3] Yung-Ju J. Chang, Gaurav Paruthi and Mark W. Newman. 2015. A Field Study Comparing Approaches to Collecting Annotated Activity Data in Real-world Settings. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 671-682. DOI: https://doi.org/10.1145/2750858.2807524

[4] Allan H. Church. 1993. Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, 57 (1). 62-79. DOI: https://doi.org/10.1086/269355

[5] Catherine E. Connelly, David Zweig, Jane Webster and John P. Trougakos. 2012. Knowledge hiding in organizations. *Journal of Organizational Behavior*, 33 (1). 64-88. DOI: https://doi.org/10.1002/job.737

[6] Tamlin Conner Christensen, Lisa Feldman Barrett, Eliza Bliss-Moreau, Kirsten Lebo and Cynthia Kaschub. 2003. A Practical Guide to Experience-Sampling Procedures. *Journal of Happiness Studies*, 4 (1). 53-78. DOI: https://doi.org/10.1023/A:1023609306024

[7] Sunny Consolvo and Miriam Walker. 2003. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing*, 2 (2). 24-31. DOI: https://doi.org/10.1109/MPRV.2003.1203750

[8] Mihaly Csikszentmihalyi, Reed Larson and Suzanne Prescott. 1977. The Ecology of Adolescent Activity and Experience. *Journal of Youth and Adolescence*, 6 (3). 281-294. DOI: https://doi.org/10.1007/BF02138940

[9] Brian Cugelman. 2013. Gamification: What It Is and Why It Matters to Digital Health Behavior Change Developers. *JMIR Serious Games*, 1 (1). DOI: https://doi.org/10.2196/games.3139

[10] Rodrigo de Oliveira, Mauro Cherubini and Nuria Oliver. 2010. MoviPill: Improving Medication Compliance for Elders Using a Mobile Persuasive Social Game. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ACM, 251-260. DOI: https://doi.org/10.1145/1864349.1864371

[11] Sebastian Deterding, Dan Dixon, Rilla Khaled and Lennart Nacke. 2011. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9-15. DOI: https://doi.org/10.1145/2181037.2181040

[12] Dominic Upton, Penney Upton and Charlotte Taylor. 2013. Increasing children's lunchtime consumption of fruit and vegetables: an evaluation of the Food Dudes programme. *Public Health Nutrition*, 16 (6). 1066-1072. DOI: https://doi.org/10.1017/S1368980012004612

[13] Rosta Farzan, Joan M. DiMicco, David R. Millen, Casey Dugan, Werner Geyer and Elizabeth A. Brownholtz. 2008. Results from Deploying a Participation Incentive Mechanism Within the Enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 563-572. DOI: https://doi.org/10.1145/1357054.1357145

[14] Joel E. Fischer, Chris Greenhalgh and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity As Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, 181-190. DOI: https://doi.org/10.1145/2037373.2037402

[15] Zachary Fitz-Walter, Dian Tjondronegoro and Peta Wyeth. 2011. Orientation Passport: Using Gamification to Engage University Students. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, ACM, 122-125. DOI: https://doi.org/10.1145/2071536.2071554

[16] J. Goncalves, S. Hosio, D. Ferreira and V. Kostakos. 2014. Game of Words: Tagging Places through Crowdsourcing on Public Displays. In *Designing Interactive Systems*, 705-714. DOI: https://doi.org/10.1145/2598510.2598514

[17] J. Goncalves, S. Hosio, J. Rogstadius, E. Karapanos and V. Kostakos. 2015. Motivating Participation and Improving Quality of Contribution in Ubiquitous Crowdsourcing. *Computer Networks*, 90. 34-48. DOI: https://doi.org/10.1016/j.comnet.2015.07.002

[18] J. Goncalves, V. Kostakos, S. Hosio, E. Karapanos and O. Lyra. 2013. IncluCity: Using Contextual Cues to Raise Awareness on Environmental Accessibility. In *International ACM SIGACCESS Conference on Computers and Accessibility*, 17:11-17:18. DOI: https://doi.org/10.1145/2513383.2517030

[19] J. Goncalves, V. Kostakos, E. Karapanos, M. Barreto, T. Camacho, A. Tomasic and J. Zimmerman. 2014. Citizen Motivation on the Go: The Role of Psychological Empowerment. *Interacting with Computers*, 26 (3). 196-207. DOI: https://doi.org/10.1093/iwc/iwt035

[20] Juho Hamari, Jonna Koivisto and Harri Sarsa. 2014. Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. In *47th Hawaii International Conference on System Sciences*, IEEE, 3025-3034. DOI: https://doi.org/10.1109/HICSS.2014.377

[21] Walt Haney, Mike Russell, Cengiz Gulek and Ed Fierros. 1998. Drawing on Education: Using Student Drawings To Promote Middle School Improvement. *Schools in the Middle*, 7 (3). 38-43.

[22] Joel M. Hektner, Jennifer A. Schmidt and Mihaly Csikszentmihalyi. 2007. *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage.

[23] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen and Deborah Estrin. 2010. AndWellness: An Open Mobile System for Activity and Experience Sampling. In *Wireless Health 2010*, ACM, 34-43. DOI: https://doi.org/10.1145/1921081.1921087

[24] Stefan E. Hormuth. 1986. The sampling of experiences in situ. *Journal of Personality*. DOI: https://doi.org/10.1111/j.1467-6494.1986.tb00395.x

[25] S. Hosio, J. Goncalves, V. Lehdonvirta, D. Ferreira and V. Kostakos. 2014. Situated Crowdsourcing Using a Market Model. In *User Interface Software and Technology*, 55-64. DOI: https://doi.org/10.1145/2642918.2647362

[26] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi and Scott E. Hudson. 2008. Using Visualizations to Increase Compliance in Experience Sampling. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, ACM, 164-167. DOI: https://doi.org/10.1145/1409635.1409657

[27] Brooke A. Jones, Gregory J. Madden and Heidi J. Wengreen. 2014. The FIT Game: preliminary evaluation of a gamification approach to increasing fruit and vegetable consumption in school. *Preventive Medicine*, 68. 76-79. DOI: https://doi.org/10.1016/j.ypmed.2014.04.015

[28] J. H. Jung, Christoph Schneider and Joseph Valacich. 2010. Enhancing the Motivational Affordance of Information Systems: The Effects of Real-Time Performance Feedback and Goal Setting in Group Collaboration Environments. *Manage. Sci.*, 56 (4). 724-742. DOI: https://doi.org/10.1287/mnsc.1090.1129

[29] Nicolas Kaufmann, Thimo Schulze and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In *Amcis*.

[30] Vassilis-Javed J. Khan, Panos Markopoulos, Berry Eggen, Wijnand Ijsselsteijn and Boris de Ruyter. 2008. Reconexp: A Way to Reduce the Data Loss of the Experiencing Sampling Method. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, 471-476. DOI: https://doi.org/10.1145/1409240.1409316

[31] Bohyun Kim. 2015. *Understanding Gamification*. American Library Association.

[32] Robert E. Kraut, Paul Resnick, Sara Kiesler, Yuqing Ren, Yan Chen, Moira Burke, Niki Kittur, John Riedl and Joseph Konstan. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

[33] Reed Larson and Mihaly Csikszentmihalyi. 1983. The Experience Sampling Method. In *Flow and the Foundations of Positive Psychology*, Wiley Jossey-Bass, 41-56.

[34] S. Lui, K. Lang and S. Kwok. 2002. Participation Incentive Mechanisms in Peer-to-Peer Subscription Systems. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society.

[35] Peter Lynn. 2001. The Impact of Incentives on Response Rates to Personal Interview Surveys: Role and Perceptions of Interviewers. *International Journal of Public Opinion Research*, 13 (3). 326-336. DOI: https://doi.org/10.1093/ijpor/13.3.326

[36] Marek Machnik, Michael Riegler and Sagar Sen. 2015. Crowdpinion: Motivating People to Share Their Momentary Opinion. In *Second International Workshop on Gamification for Information Retrieval*.

[37] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic and Mirco Musolesi. 2015. Ask, But Don't Interrupt: The Case for Interruptibility-Aware Mobile Experience Sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ACM, 723-732. DOI: https://doi.org/10.1145/2800835.2804397

[38] B. Morschheuser, J. Hamari and J. Koivisto. Year. Gamification in Crowdsourcing: A Review. In *49th Hawaii International Conference on System Sciences*, Hawaii, 4375-4384. DOI: https://doi.org/10.1109/HICSS.2016.543

[39] Mohamed Musthag, Andrew Raij, Deepak Ganesan, Santosh Kumar and Saul Shiffman. 2011. Exploring Micro-Incentive Strategies for Participant Compensation in High-burden Studies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, ACM, 435-444. DOI: https://doi.org/10.1145/2030112.2030170

[40] Leysia Palen and Marilyn Salzman. 2002. Voice-mail Diary Studies for Naturalistic Data Capture Under Mobile Conditions. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, ACM, 87-95. DOI: https://doi.org/10.1145/587078.587092

[41] Shyam Reyal, Shumin Zhai and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 679-688. DOI: https://doi.org/10.1145/2702123.2702597

[42] Chad Richards, Craig W. Thompson and Nicholas Graham. 2014. Beyond Designing for Motivation: The Importance of Context in Gamification. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, ACM, 217-226. DOI: https://doi.org/10.1145/2658537.2658683

[43] Christie N. Scollon, Chu Kim-Prieto and Ed Diener. 2003. Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies*, 4 (1). 5-34. DOI: https://doi.org/10.1023/A:1023605205115

[44] Arthur A. Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, 16 (3). 199-202.

[45] Arthur Stone, Ronald Kessler and Jennifer Haythomthwatte. 1991. Measuring Daily Events and Experiences: Decisions for the Researcher. *Journal of Personality*, 59 (3). 575-607. DOI: https://doi.org/10.1111/j.1467-6494.1991.tb00260.x

[46] Yoshitaka Ueyama, Morihiko Tamai, Yutaka Arakawa and Keiichi Yasumoto. 2014. Gamification-Based Incentive Mechanism for Participatory Sensing. In *IEEE International Conference on Pervasive Computing and Communications Workshops*, IEEE, 98-103. DOI: https://doi.org/10.1109/PerComW.2014.6815172

[47] Niels van Berkel, Chu Luo, Theodoros Anagnostopoulos, Denzil Ferreira, Jorge Goncalves, Simo Hosio and Vassilis Kostakos. 2016. A Systematic Assessment of Smartphone Usage Gaps. In *Proceedings of the Conference on Human Factors in Computing Systems*, 4711-4721. DOI: https://doi.org/10.1145/2858036.2858348

[48] Aku Visuri, Niels van Berkel, Jorge Goncalves, Chu Luo, Denzil Ferreira and Vassilis Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In *Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services*, to appear. DOI: https://doi.org/10.1145/3098279.3098532

[49] Luis von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51 (8). 58-67. DOI: https://doi.org/10.1145/1378704.1378719

[50] Zhen Yue, Eden Litt, Carrie J. Cai, Jeff Stern, Kathy Baxter, Zhiwei Guan, Nikhil Sharma and Guangqiang Zhang. 2014. Photographing Information Needs: The Role of Photos in Experience Sampling Method-style Research. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 1545-1554. DOI: https://doi.org/10.1145/2556288.2557192