

SpaceMaze: Incentivizing Correct Mobile Crowdsourced Sensing Behavior with a Sensified Minigame

Matthias Budde^a, Jan Felix Rohe^a, Lina Hirschhoff^a, Patrick Schlosser^a, Michael Beigl^a, Jussi Holopainen^b, and Andrea Schankin^a

^aKarlsruhe Institute of Technology (KIT), TECO / Pervasive Computing, Vincenz-Priessnitz-Strasse 1, Karlsruhe, Germany; ^bUniversity of Lincoln, Brayford Pool, Lincoln, United Kingdom

ABSTRACT

Modern mobile phones are equipped with many sensors, which can increasingly be used to sense various environmental phenomena. In particular, mobile sensing has enabled crowdsourced data collection at an unprecedented scale. However, as laypersons are involved in this, concerns regarding the data quality arise.

This work explores the gamification of smartphone-based measurement processes in practice by embedding a sensing task into a mobile minigame. The underlying idea is — rather than to educate the user on how to correctly perform a measurement task — to opportunistically execute the measurement in the background once the smartphone is in a suitable context.

To this end, this paper presents the design and evaluation of SpaceMaze, a smartphone game with the goal of minimizing user error by introducing appropriate game mechanics to influence the phone context, using the example of mobile noise level monitoring.

A large user study that compares SpaceMaze to two non-gamified apps for noise level monitoring (N=360 in total) shows that SpaceMaze can successfully reduce user errors when compared to simple non-gamified ambient noise level monitoring applications and that the minigame is generally perceived as being enjoyable. Solutions for remaining problems, such as noise generated by the players, are discussed.

KEYWORDS

human error; data quality; ubiquitous computing; gamification; human-machine interface; environmental sensing

1. Introduction

Sensors, connectivity and computing power of today’s smartphones make them an ideal platform for so-called Participatory Sensing (Burke, Estrin, Hansen, Parker, Ramathan, Reddy, and Srivastava, 2006). In Participatory Sensing, citizens and groups, especially non-experts, perform data collection and interpretation tasks in the environments they live in, using widely available and familiar technology, particularly their own personal devices. Participatory sensing projects range from plant observation (Han, Graham, Vassallo, and Estrin, 2011) to monitoring environmental phenomena such as noise pollution (Maisonneuve, Stevens, and Ochab, 2010).

While such systems have a large potential, participatory sensing challenges researchers in at least two aspects. First, potential users need to be engaged in the task. One approach used in different application areas to facilitate user engagement is

the implementation of gamification strategies, i.e. "the use of game design elements in non-game contexts" (Deterding, Dixon, Khaled, and Nacke, 2011). The idea of gamification is to derive the fun and addicting elements of games in order to motivate people performing a task that might not be engaging enough by itself. However, gamification strategies are not undisputed in participatory sensing. It has been argued that they may have negative effects on intrinsic motivation of people who are already highly engaged in participatory sensing (Mekler, Brühlmann, Opwis, and Tuch, 2013). A second issue of participatory sensing is that the quality of the collected data may be affected by many diverse sources of systematic error (Budde, Köpke, and Beigl, 2015; van Berkel, Budde, Wijenayake, and Goncalves, 2018). This leads to opportunities to use crowdsourced data collection to achieve positive outcomes for science and society going unrealized and even existing data being disregarded because of doubts concerning its validity (Harding, Knowles, Davies, and Rouncefield, 2015; Gibb, 2015). However, if properly familiarized with the task, research has shown that laypersons can collect data of comparable quality to experts (See, Comber, Salk, Fritz, van der Velde, Perger, Schill, McCallum, Kraxner, and Obersteiner, 2013).

This work explores the using game elements in a smartphone-based measurement process by embedding sensing tasks into a mobile minigame. We apply the concept of *Sensified Gaming* (Budde, Öxler, Beigl, and Holopainen, 2016), which, in contrast to adding game elements to a task, proposes to embed sensing tasks into a full-fledged game. The underlying idea is – rather than to educate or to motivate the user on how to correctly perform a measurement task – to opportunistically execute the measurement in the background once the phone is in a suitable context.

The mobile minigame is designed in a way that supports automatic detection of the suitable context and ensures that just by playing the game the user goes through a correct measurement process. We would like to note that this approach is not intended as a general solution to crowdsourced data collection, but rather as a complementary method. Especially in projects of a Citizen Science nature, gamification approaches can also have adverse effects on engagement, particularly on intrinsically motivated people (Mekler et al., 2013). A deeper discussion on drawbacks of a game-based approach is included in the discussion section of this paper.

In this work, we present the design and evaluation of *SpaceMaze*, a sensified game for noise level monitoring. The discussion of the identified game design features and the evaluation of their suitability for supporting correct sensing behaviour form the main contribution of this study. We also include discussions on possible ethical aspects, the design effort associated with the approach and implications of somewhat less than perfect choice of game mechanics, which unfortunately, was the case in one aspect of our specific game design. However, as a result, our discussion also incorporates possible mitigation approaches in which parts of the collected data are selectively discarded based on in-game events. The motivational aspects of the game in regards to increasing the likelihood to first start and then continue contributing to a participatory sensing project are briefly discussed as well.

2. Related Work

A range of approaches has been presented in the past to increase the quality of data collection tasks, from training (Sheppard and Terveen, 2011) through selecting users (Wickens, Lee, Liu, and Gordon-Becker, 2014) to providing instructions to using sensors to verify the sensing context. While the latter approach has shown the capability

to reduce the amount of errors that users make, problems remain, e.g. users might perceive the task as unpleasant, not understand it, etc. Important aspects are also, that many of these approaches do not scale to Participatory Sensing scenarios in which a large crowd of users is involved in the data collection (Budde, Schankin, Hoffmann, Danz, Riedel, and Beigl, 2017a), and that most of these mechanisms are only applicable when users are already engaged.

Engaging users in Participatory Sensing in the first place was extensively studied by Restuccia, Das, and Payton (2016). Their survey of incentives for applications where participants voluntarily contribute sensor data presents a taxonomy of incentive mechanisms, discusses them in-depth and concludes with an agenda of open research challenges. They argue that past research often assumes that all people have the same needs and call for application-tailored and/or behavior-conscious incentive mechanisms, away from one-size-fits-all solutions. They also recognize that “*one of the greatest limitations of existing incentive mechanisms, and more generally, of participatory sensing systems, is that they are not effective in tackling the problem of verifying the actual QoI [Quality of Information] of sensing reports*” (Restuccia et al., 2016).

Gamification and Serious Games (or Games with a purpose (GWAPs) (Lafourcade, Joubert, and Le Brun, 2015; Von Ahn, 2006)) have been used as motivational tools to attract new volunteers and audiences for citizen science campaigns (Bowser, Hansen, He, Boston, Reid, Gunnell, and Preece, 2013) as well as to further engage existing users and sustain their engagement (Crowston and Prestopnik, 2013; Iacovides, Jennett, Cornish-Trestrail, and Cox, 2013). Morschheuser, Hamari, Koivisto, and Maedche (2017) presented a comprehensive overview of gamified crowdsourcing approaches, the focus again lying on motivation. An example of a system gamifying a data collection task is *Forgotten Island*, a game created with the purpose of motivating players to participate in the categorization of moths. It showcases that players can be engaged by a real, full-fledged game and at the same time still help with the evaluation of research data (Crowston and Prestopnik, 2013). However, it does not involve mobile devices or the collection of sensor data, which are some of the defining characteristics of Participatory Sensing scenarios, as the one underlying *SpaceMaze*.

Simultaneously looking at the quality of the collected data has received less attention (Budde et al., 2017a; Simperl, Reeves, Phethean, Lynes, and Tinati, 2018). Prestopnik, Crowston, and Wang (2014) showed that games with a purpose can in principle yield contributions of equal quality. In terms of increasing data quality, the location-based game *GeoSnake* (Matyas, Kiefer, Schlieder, and Kleyer, 2011) has been used to boost verification rates, an approach which can only be applied to data collection tasks that are repeatable or reviewable later, which is not the case for sensor data collection. The same applies to the gamification approaches in the context of Experience Sampling presented by Van Berkel, Goncalves, Hosio, and Kostakos (2017), who demonstrated that gamification techniques can increase data quality and quantity. *TheBabyLaunch* is an example of a calibration minigame in which a previously unexciting sensor data collection task was gamified and eventually perceived as much more enjoyable than the standard system (Flatla, Gutwin, Nacke, Bateman, and Mandryk, 2011). It is used to calibrate a physical sensor that measures chest inhalation and adds a gamified visualization to the process. Unfortunately, the approach can not directly be transferred to the recording of uncontrolled and/or non-user-generated sensor data, such as in environmental sensing used in *SpaceMaze*. *PhotoCity* (Tuite, Snavely, Hsiao, Tabing, and Popovic, 2011), on the other hand, gamifies the training of taking urban photographs suitable for creating 3D building models for online mapping tools. This approach can help against mistakes (errors in the intention) but cannot reduce slips

(errors in the execution, cmp. Norman (1983)) such as inadvertently failing to follow a correct measurement process.

Sensified Gaming (Budde et al., 2016) has been proposed as an approach that embeds the sensing task into a game rather than gamifying the task itself, the idea being to align gameplay and desired real-world behavior. A similar approach was brought forth by Kayali, Luckner, Purgathofer, Spiel, and Fitzpatrick (2018) in the domain of games for health: “*In games for health there is ideally an overlap of core-gameplay with health-beneficial behavior. In contrast to gamification techniques, [...], games for health can use game mechanics that are aligned with the behavior they are meant to support*” (Kayali et al., 2018). Budde et al. (2016) provide a review and classification of game design patterns (Björk and Holopainen, 2004; Björk, 2009) suitable for identifying or designing game contexts that are used to ensure the correct execution of sensor data collection tasks. This is also the approach underlying *SpaceMaze*.

3. Game Design

SpaceMaze was developed following the concept of *Sensified Gaming* (Budde et al., 2016), i.e. we identified the constraints and errors that users are likely to make in the sensing task and then chose appropriate game design patterns so that the game would guide the user to not making said errors.

We chose the scenario of *Participatory Noise Level Monitoring*, as it is a somewhat intuitive task and can be easily be done with off-the-shelf mobile phones. It also has the benefit that it is a well-understood Participatory Sensing task that has been realized in the past by a number of authors (Maisonneuve et al., 2010; Rana, Chou, Kanhere, Bulusu, and Hu, 2010; Kanjo, 2010; Schweizer, Bärtl, Schulz, Probst, and Mühläuser, 2011; Muratori, Salomoni, and Pau, 2011) and has been analyzed before in terms of data quality respectively user error (Budde et al., 2017a). This gives us a baseline to compare the game against in our evaluation.

The idea behind *SpaceMaze* is simple: The phone’s internal sensors can be used to determine whether the sensing context is correct and allow or deny the recording of ambient audio levels accordingly. Combining this with a game design pattern based design approach results in a game in which the recording of an environmental phenomenon can opportunistically be carried out in the background.

3.1. Constraints

The focus of this work is to find ways to minimize the variance in exhibited behavior when measuring, specifically the kind that may have an adverse effect on data quality. We refer to this kind of behavior as *human error* in this paper, following Norman (1983): In this sense, human error both covers *mistakes* (errors in the intention) and *slips* (errors in carrying out the intention). We also explicitly include “mistakes” that are beyond the direct control of the user or caused because the user had incorrect or incomplete information on the task. This does not imply that the user is to blame. Still, according to the definition above – and in terms of the negative effect on data quality – they remain errors.

As requirements of the sensing scenario, we have collected handling and behavior constraints from the best practices for noise level monitoring from the PDF user guide of the *NoiseTube* project (Maisonneuve et al., 2010). A deviation from this behavior while collecting data would be considered an error (either slip or mistake).

When recording, ...

- the device should be rotated by 180 degrees, so that the microphone points away from the user.
- the device should be held very still, no shaking or moving too much.
- a recording must at least 30 seconds long to be valid.
- the user should not talk or make other noise.
- the user should be outside.
- no phone usage, e.g. typing, chatting or any other device functions for the duration of the recording.

The latter two aspects are actually hard to control. Ideally, the game would somehow encourage the player not to talk or create additional noise during the gameplay session. However, since the microphone of the device is already recording the noise levels, it is not able to also interrupt measurement when the user is making noise. On a purely technical level, it is not possible to ensure that the player is quiet during the game. We theorized that challenging gameplay might help to discourage the user to make unwanted noise, as would the knowledge that *SpaceMaze* is a sensing game that performs measurements in the background.

3.2. Genre and Game Mechanics

A game design pattern based approach was followed when building *SpaceMaze* (Budde, 2018). This means that the design of the game, from the type and genre down to the specific game mechanics, was built using a collection of documented solutions to recurring design challenges (Björk, 2009; Björk, Lundgren, and Holopainen, 2003). The choice of the particular patterns was strongly driven by the constraints that were presented in the previous section. These constraints exclude certain types of games right from the start.

The game in question needs to be a relatively short game (in terms of playing time for one game instance) that can be played outside in a potentially noisy environment. This excludes otherwise interesting genre choices such as meditation games, as it is unlikely that players would enjoy attempting to play such games e.g. in the vicinity of a noisy crossroads. This poses a dilemma: Since the phenomenon that is sensed in our use case is ambient noise, we cannot make sound an input for the game. Otherwise, the (desired) measurement of a noisy environment could have unintended effects on the game. On the other hand, this makes it impossible to use game elements to directly keep the player from making noise, for instance by “punishing” player generated noise. Therefore we chose a somewhat demanding game to hopefully inhibit player generated noise. Secondly, the game must vary to prevent it from becoming boring when played over and over again. Therefore, to reduce the risk of churn, we chose a game with a level-based approach that allows automatic generation of levels: a maze-like level without any dead ends. As a result, *SpaceMaze* features several different visible game elements (see Figure 1) that correspond to an appropriate game design pattern, as explained below.

In *SpaceMaze*, the players need to steer a spaceship through a labyrinth (see Figure 2). The ship always moves and cannot be stopped. Yellow orbs are placed throughout the level as pickups. When collected, they increase the ship’s movement speed, potentially allowing the player to clear the level faster.



Figure 1. Game elements of *SpaceMaze* (left-to-right): green finish line, light green checkpoint, blue moving obstacle, yellow pickup, player space ship with surrounding shield, red player-killing obstacle, and warp point.



Figure 2. Screenshots from the gameplay of *SpaceMaze*.

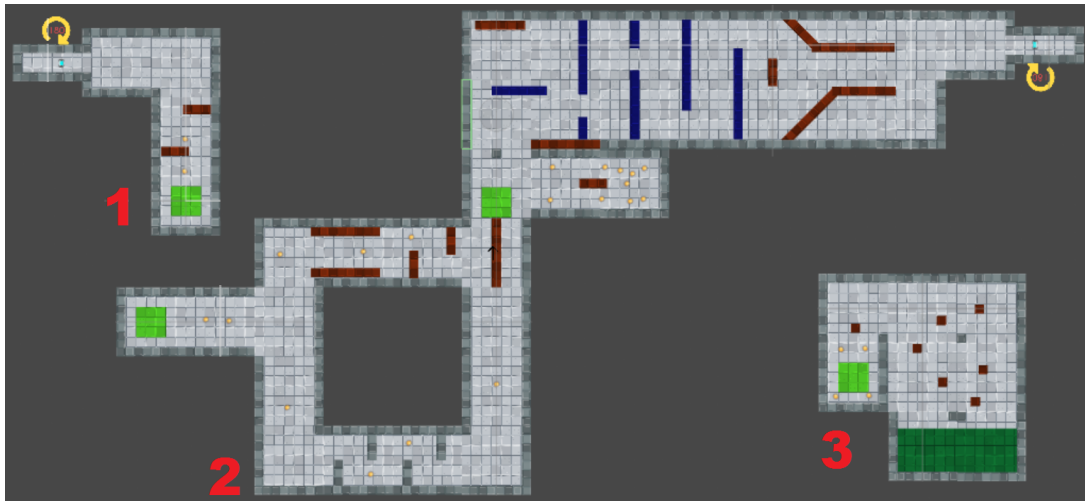

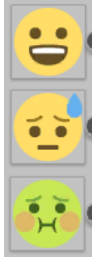




Figure 3. Map of the first level of *SpaceMaze*: Players start in part 1 and are beamed to part 2 using a warppoint. This larger area is where the phone opportunistically performs the ambient audio sensing in the background. After having cleared the obstacle course the ship is eventually warped to part 3 of the level, leaving the user with a normally oriented phone on level completion.

Table 1. Selected game design patterns and their effect concerning the constraints of the measurement task.

Task/ Constraint	Design Pattern(s)	Game Element	
Avoid shaking or movement; no phone usage	<i>Mimetic interfaces;</i> <i>Player physical prowess</i>	The spaceship is controlled by rotating the device ever so slightly. A transmission of 1:12 makes sure that the phone movement is minimized: 1° of rotation of the phone causes 12° of rotation in-game.	
Avoid shaking or movement; no phone usage	<i>Mimetic interfaces;</i> <i>Game items that protect from damage</i>	In the game’s status bar, the status of the ship’s engineer is shown. If the device is shaken or held in the wrong orientation, the engineer get’s sick and the ship’s shields power down. Without shields, the ship will explode when touching anything, e.g. the level walls.	
Device should be held upside-down	<i>Mimetic interfaces</i>	Warp-points transport the ship from one part of the level to another. To activate them, the user needs to turn the phone by 180°.	
Record for at least 30 s	<i>Real-time games;</i>	The ship is always moving forward and can not be stopped. The camera is locked on the ship, keeping it centered in the display. The continuous gameplay facilitates longer periods in which recordings can be made.	
Record for at least 30 s	<i>Pre-defined goals</i>	The levels have a linear structure, i.e. the user knows a-priori that he simply needs to play until reaching the end of the level, giving ample time for recording in the background.	

A shield protects the ship from damage, preventing it from being destroyed on collisions with blue obstacles or the level walls. Red obstacles will always destroy the ship, regardless of its shield’s status. Upon destruction, the ship is respawned instantaneously at the last checkpoint. The total time upon finishing a level directly determines the achieved points.

Aside from the general gameplay, some of the elements were specifically selected to guide the device handling during the measurement process (see Table 1): The most central design pattern is *Mimetic Interfaces* (Björk, 2009). By controlling the ship via the phone’s inertial sensors (accelerometer and gyroscope), well-defined constraints are placed on the possible device motion.

Since one of the measurement constraints is that the device must be held relatively steady, we designed the controls using a transmission ratio of 1:12. This means that for every degree that the phone is rotated, the spaceship in the game rotates by 12 degrees. While in the real world, the smartphone is only turned slightly, in-game, the ship rotates much more, forcing the player to turn the device very carefully and

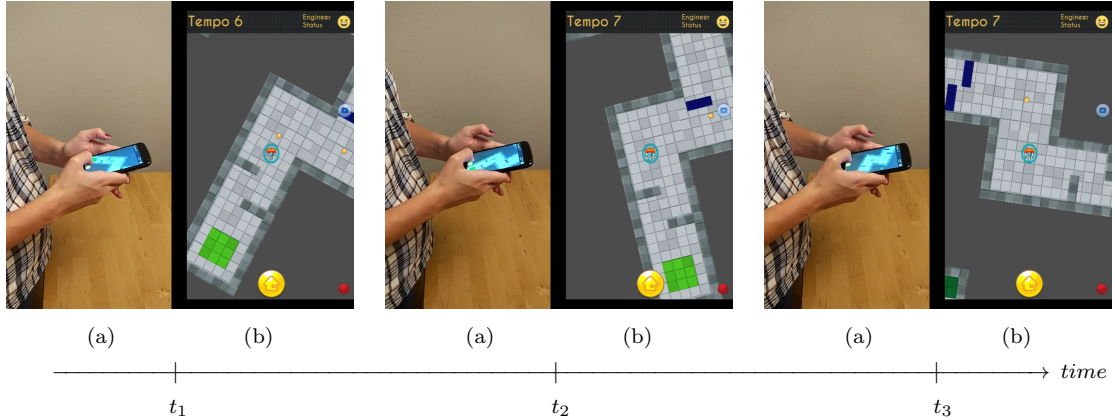


Figure 4. *SpaceMaze* is controlled using a mimetic interface using a transmission ratio of $1:12$. While in the real world (a), the smartphone is only turned slightly, in-game (b), the ship rotates more than 45° , resulting in a quasi-motionless device handling.

resulting in a quasi-motionless device handling (see Figure 4).

Mimetic interfaces are again used in conjunction with the *warp points* game element. In order to beam the spaceship to the next stage of the level, the user needs to rotate the phone by 180° . By that, the correct orientation of the phone for a measurement (i.e. microphone pointing away from user) is ensured. Design patterns used to ensure a certain length of recordings are *Real-time games* and *Pre-defined goals*, as it takes a certain time to clear a level.

3.3. Level Design

The main level of *SpaceMaze* is split into three parts (see Figure 3). The first and the third part of the level are not relevant for the sensing task but only to initially establish the correct phone context. Players start in part 1 and are beamed to part 2 using a warppoint. This makes sure that the user has turned the phone by 180° .

The larger area that follows is where the phone opportunistically performs the ambient audio sensing in the background. It is composed of an obstacle course through which the spaceship must be navigated. After having cleared the maze, the ship is eventually warped to part 3 of the level, leaving the user with a normally oriented phone on level completion.

4. Evaluation

We conducted a user study in order to explore whether the design of *SpaceMaze* indeed has a positive effect on the performance of the measurement process, and if so, how this relates to two of our non-gamified noise-level recording app prototypes, which we had developed with the same design goal (Budde et al., 2017a).

The two app flavors we chose as reference are depicted in Figure 5. The *Baseline* app flavor is a recording app that just features one animated recording button and a time display. The *Context-Sensitive* app flavor features technological measures to guide the sensing process: On startup of the app, the interface is rendered upside-down, so that the user would turn around the phone. The phone's sensors are used to monitor device movement and check if the phone receives a GPS signal. The recording button is only

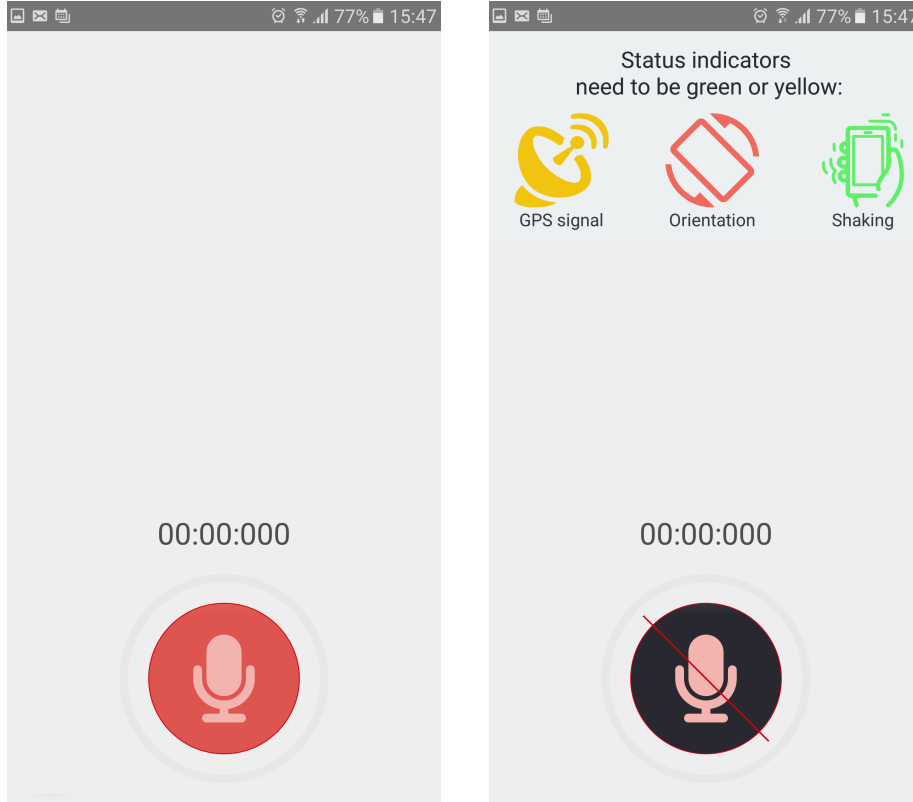


Figure 5. Non-gamified apps (from Budde et al. (2017a)): The *Baseline* app flavor (left) is just a simple one-button recording application and the *Context-Sensitive* flavor features technological measures to guide the sensing process: The interface is upside-down, sensors are used to monitor device movement, and the recording button is only enabled when the constraints are met.

enabled when the constraints are met.

4.1. Study Design

We compare *SpaceMaze* against the two other noise-level recording apps in a between-subject design. Each group only used one of the three app prototypes in order to control for learning effects about the correct measurement process. Except for the app tested, the study protocol was the same.

4.2. Participants

In total 360 participants took part in the study. The study was carried out in accordance with the ethical standards (i.a. informed consent, privacy considerations, etc.) laid down in the 1964 Declaration of Helsinki and its amendments (World Medical Association, 1964). For the evaluation of *SpaceMaze*, we recruited 107 study participants (82 male, 24 female, 1 not specified). The evaluation of the apps we compared *SpaceMaze* against had been carried out with a similar user group and sample size ($N_{Baseline}=123$, $N_{Context-Sensitive}=130$). Generally, the participant groups were of a comparable social profile. That means that the sample was heterogeneous with regard to age and education, with a tendency towards a technical background. All participants

were well familiar with using mobile phones, most of them also using a smartphone. None of the participants were experts in Participatory Sensing, only few of them had participated in data collection campaign before.

In the following, we describe the *SpaceMaze* group in more detail. Detailed information about the other two groups are reported in Budde et al. (2017a).

The study was conducted face-to-face in field: We approached passing pedestrians inside a city and on a university campus and asked them if they would volunteer to participate (sample of opportunity, response rate at ca. 2/3). Almost all participants were accustomed to mobile phone usage, stating they used their device daily (N=104). Additionally, close to half regularly played games on their mobile device (N=46). All participants gave their written informed consent and were compensated with a gift certificate with a value of 5 Euro each for their participation. Most of the participants had a technical, natural science or IT background (N=72), followed by an economic background (N=14). The remaining 18 participants were evenly distributed over medicine, law, education, arts, administration, craft and others. In total, we collected an educated group of participants, most having a technical background, being familiar with smartphones and partially having experience with games on a regular basis.

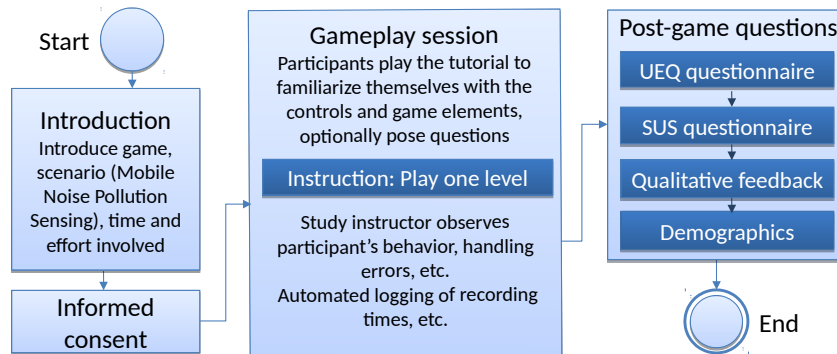


Figure 6. Session structure

4.3. Session Structure

After having introduced the scenario, participants played the tutorial level once to familiarize themselves with the principle of the game. Then, they were asked to play one full level until they completed it or gave up. No further instructions were given.

The study instructor observed participants and recorded all user errors. In addition, the game recorded automatically logged gameplay data. No actual audio recordings were made by the device. After having finished, participants were asked to fill in the standard *User Experience Questionnaire (UEQ)* (Laugwitz, Held, and Schrepp, 2008), the *System Usability Scale (SUS)* (Brooke, 1996) and provide written qualitative feedback.

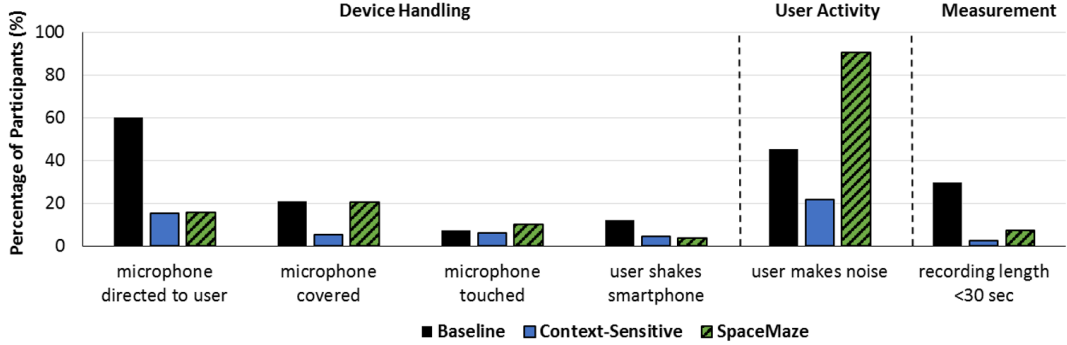


Figure 7. Measurement errors observed during the sensing process in the mobile minigame *SpaceMaze* in comparison to the non-gamified applications *Baseline* and *Context-Sensitive* (Budde et al., 2017a).

5. Results

To assess the suitability of *SpaceMaze* for guiding the user to perform a correct measuring process, we compared our data with that of two non-gamified applications, i.e., *Baseline* and *Context-Sensitive* as reported in previous research (Budde et al., 2017a).

5.1. Quantitative Analysis of the Influence of Game Elements on User Behavior

To assess whether game elements guide the user in the measuring process, we observed the user behavior in three main categories, i.e. device handling, user activity, and measurement duration (retrieved from the game log). We report the percentage of participants who showed incorrect behavior in each category compared to the two non-gamified apps *Baseline* and *SpaceMaze*. For statistical validation, we run 2 by 2 Chi-Square tests with the variables app (*SpaceMaze* vs. either *Baseline* or *Context-Sensitive*) and behavior (correct vs. incorrect) for each category. As we compare *SpaceMaze* with both *Baseline* and *Context-Sensitive*, we corrected the level of significance according to Bonferroni, that is effects are statistically significant if the observed $p < 0.025$. To assess the practical meaning of the empirical results, we calculated ϕ as effect size. Following the conventions by Cohen (Cohen, 1988), the effect is small if $0.10 < \phi < 0.30$, medium if $0.30 < \phi < 0.50$, and large if $\phi > 0.50$.

Figure 7 displays the observed user errors of all three applications. Refer to Table 2 for an overview of the mean, statistical significance, and effect sizes.

In comparison to the *Baseline* app, we observed a significant drop in errors. Only 16% of the participants still directed the microphone to themselves. These were about 44% less than with the *Baseline* app ($\chi^2(1) = 46.911$, $p < 0.001$, $\phi = 0.45$). About 4% of the participants still shook their smartphone, i.e., 9% less than with the *Baseline* app ($\chi^2(1) = 5.400$, $p = 0.020$, $\phi = 0.45$). Compared to the *Context-Sensitive* app, we observed no significant improvement in any error category. Instead, the microphone was covered by 15% more participants when using *SpaceMaze* than the *Context-Sensitive* app ($\chi^2(1) = 12.429$, $p < 0.001$, $\phi = 0.23$). About 92% of the participants provided the minimal recording length of 30 seconds, i.e., 22% more than with the *Baseline* app ($\chi^2(1) = 18.140$, $p < 0.001$, $\phi = 0.29$). The errors "microphone covered" and "microphone touched" were observed equally often ($\chi^2(1) = 0.012$, $p = 0.914$, $\phi = 0.01$ or

Table 2. Overview of user error values

Variable	<i>SpaceMaze</i>	<i>Baseline</i>		<i>Context-Sensitive</i>	
	% Part.	% Part.	ϕ	% Part.	ϕ
Microphone directed to user	15.9	60.2***	0.45	15.5	0.01
Microphone covered	20.6	21.1	0.01	5.4***	0.23
Microphone touched	10.3	7.3	0.05	6.2	0.08
User shakes smartphone	3.7	12.2*	0.15	4.7	0.02
Recording length <30 sec	7.5	29.9***	0.29	2.5	0.12
User makes noise	90.7	45.5***	0.48	21.7***	0.69

Note: * $p < 0.025$, *** $p < 0.001$

Table 3. Noise categories of the participants' behavior while playing *SpaceMaze*

Noise category	number of participants
Speaking, whistling, voice sounds	79
Walking, step sounds	4
Tapping, hitting device	7
Scratching, clapping, knocking sounds	5
Breathing, sneezing, coughing sounds	50
Frustration noise	40
Game-related statements	55

$\chi^2(1) = 0.633$, $p = 426$, $\phi = 0.05$, respectively).

In comparison to both the *Baseline* and *Context-Sensitive* app, there was an outstanding increase in the category "user makes noise" ($\chi^2(1) = 52.322$, $p < 0.001$, $\phi = 0.48$ or $\chi^2(1) = 111.611$, $p < 0.001$, $\phi = 0.69$, respectively). Analyzing the types of noise in more detail, 55 participants made game-related statements while playing (for example "This game is difficult."), 40 participants made noise out of frustration over the game. Other sources of user noise were speaking in general, walking, tapping or hitting the device, coughing or sneezing. All these possible sources of noise resulted in 97 out of 107 participants in total making noise during the game. Table 3 displays the noise categories and numbers in detail. We discuss this effect in more detail below.

5.2. Quantitative Analysis of the Influence of Game Elements on Usability and User Experience

To answer the question whether the mobile minigame is more enjoyable than other applications for noise measurements by offering a better user experience, we compared *SpaceMaze* with either *Baseline* or *Context-Sensitive* with regard to the subjective ratings in the UEQ and the SUS. The reliability of these questionnaires (as estimated as internal consistency Cronbach’s α) was good for the SUS ($\alpha = 0.80$). For the UEQ, the reliability of two scales was poor (*Dependability*: $\alpha = 0.58$) or questionable (*Efficiency*: $\alpha = 0.64$). For the other scales, it ranged between acceptable (*Perspicuity*: $\alpha = 0.75$) and good (*Novelty*: $\alpha = 0.85$, *Stimulation*: $\alpha = 0.86$, and *Attractiveness*: $\alpha = 0.90$).

Because the assumption of normal distribution was violated for most scales (as indicated by Shapiro-Wilk tests), we report medians and interquartil ranges (IQR) instead of means and standard deviations. For statistical validation, we used non-parametric tests (Mann-Whitney U). As we compare *SpaceMaze* with both *Baseline* and *Context-Sensitive*, we corrected the level of significance according to Bonferroni, that is effects are statistically significant if the observed $p < 0.025$. To assess the practical meaning of the empirical results, we calculated the effect size r . Following the conventions by Cohen (Cohen, 1988), the effect is small if $0.10 < r < 0.30$, medium if $0.30 < r < 0.50$, and large if $r > 0.50$.

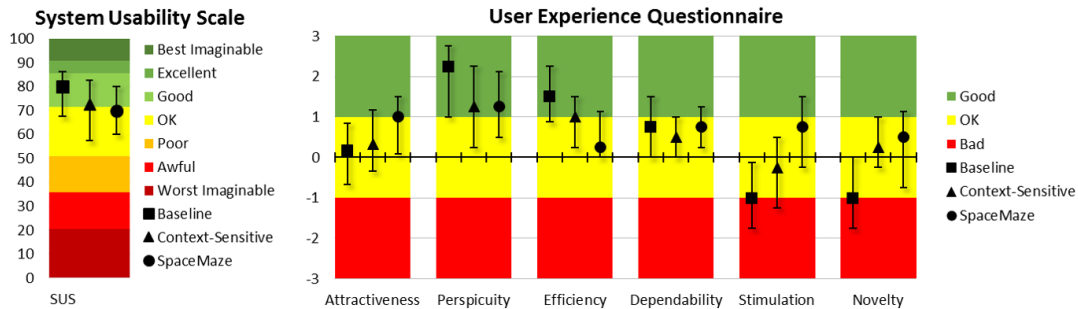


Figure 8. Results from the *System Usability Scale (SUS)* and the *User Experience Questionnaire (UEQ)*. Symbols represent medians, whiskers indicate lower and upper quartil (IQR).

Figure 8 displays the results of the two questionnaires. Compared to the *Baseline* app, we notice a significant improvement in *Attractiveness* ($U(230) = 4576.5$, $p < 0.001$, $r = 0.26$), which measures the overall impression of the apps in terms of like / dislike. Hedonic aspects of user experience as measured with the scales *Stimulation* ($U(230) = 2706.5$, $p < 0.001$, $r = 0.51$) and *Novelty* ($U(230) = 3526.5$, $p < 0.001$, $r = 0.40$) improved significantly. In contrast, pragmatic aspects of user experience (i.e., the usability) aggravated. While *Dependability* shows only a small and insignificant decrease ($U(230) = 6537.5$, $p = 0.932$, $r = 0.01$), *Perspicuity* ($U(230) = 4252.0$, $p < 0.001$, $r = 0.31$) and *Efficiency* ($U(230) = 2698.0$, $p < 0.001$, $r = 0.51$) dropped significantly from *Baseline* to *SpaceMaze*. This was also visible in a significant reduction of the SUS score by 10 points from 80 to 70 ($U(218) = 4234.0$, $p < 0.001$, $r = 0.24$).

In comparison to the *Context-Sensitive* application, we observed a slight improvement in *Attractiveness* ($U(236) = 5594.5$, $p = 0.012$, $r = 0.16$). As part of hedonic aspects of user experience, *Stimulation* improved significantly ($U(236) = 4255.0$,

Table 4. Overview of Usability / UX evaluation results.

Dimension	<i>SpaceMaze</i>	<i>Baseline</i>		<i>Context-Sensitive</i>	
	Median	Median	r	Median	r
Attractiveness	1.00	0.17***	0.26	0.33	0.16
Perspicuity	1.25	2.25***	0.31	1.25	0.03
Dependability	0.75	0.75	0.01	0.50*	0.15
Efficiency	0.25	1.50***	0.51	1.00***	0.28
Stimulation	0.75	-1.00***	0.51	-0.25***	0.33
Novelty	0.50	-1.00***	0.40	0.25	0.00
SUS	70.0	80.0***	0.24	72.5	0.06

Note: * $p < 0.025$ *** $p < 0.001$

$p < 0.001$, $r = 0.33$), whereas *Novelty* was comparable between both apps ($U(236) = 6865.5$, $p = 0.945$, $r = 0.00$). Both apps only slightly differ in pragmatic aspects of user experience. They were rated similarly in *Perspicuity* ($U(236) = 6698.5$, $p = 0.697$, $r = 0.03$). *SpaceMaze* was slightly better in *Dependability* ($U(236) = 5737.5$, $p = 0.025$, $r = 0.15$) and significantly worse in *Efficiency* ($U(236) = 4654.0$, $p < 0.001$, $r = 0.28$). This mixed picture was reflected in similar SUS scores (70 vs. 72.5) as rating of usability ($U(228) = 5989.0$, $p = 0.390$, $r = 0.06$). Refer to Table 4 for an overview of the medians and p-values of all user experience variables.

5.3. Qualitative Analysis of the Influence of Game Elements on User Experience

In addition to the UEQ and SUS scores, we evaluated the qualitative feedback the users were prompted for in the *SpaceMaze* evaluation. On our questionnaire, we asked the participants to note down three to five points they liked about the *SpaceMaze* game, and three to five points they disliked. Most participants liked how easy it is to control the spacecraft, and how the steering is explained in the tutorial. Also, the highly sensitive steering mechanism, using the z-axis of the device, was often remarked as positive. Most users liked the game graphics and also commented positively on the challenge the game provides. Among many other positive comments, the game idea and the different obstacles and barriers were also mentioned many times.

On the negative side, the steering mechanism was mentioned most often. In this regard, our participants often criticized the steering as "too sensitive", and "not precise enough". Also, the warp portals were often mentioned as a negative point. Furthermore, for 18 participants each, the game was too easy and therefore boring, the design was too simple, or they did not like that they had no direct control over the spacecraft's speed. Additionally, 12 participants mentioned that the bounding box of the spacecraft did not exactly match its visible shape, resulting in unjustified deaths, and

the same number of participants did not like the game elements of the barriers in general. A summary of the most commented about positive and negative aspects of the game can be seen in Table 5.

Table 5. Top positive and negative comments from the *SpaceMaze* evaluation (post-game, written qualitative feedback).

Positive Comments	N	Negative Comments	N
Easy to learn	47	Steering mechanism	32
Steering mechanism	32	Warp portals	26
Design, graphics	28	Too easy, boring	18
Challenge	18	Design too simple	18
Game idea	15	No direct control over speed	18
Obstacles and barriers	14	Hitbox of the aircraft	12
Level design	13	Barriers	12
Good pastime	11	Structure of levels	11
Adaptive speed	11	Unclear errors	10
Exciting/suspenseful	10	Bugs, immature software	9
Sensified gaming approach	9	Length of tutorial	9
Fun	9	Too difficult	8
Gold orbs	7	Checkpoints, re-spawning	7
Warp portals	7	Hard to learn	7
Runs smoothly	5	Engineer and shield unclear	7
Checkpoints	4	Calibration of controls	7
Story	3	No originality	6
Engineer/Shield	2	Missing high score	4
Space theme	2	Missing sound effects	2

6. Discussion

We discuss the presented results from our field study with respect to two research questions, namely whether we can guide the user towards performing a correct sensing procedure and whether the user experience increases by framing the measurement process as a game.

In our field study, we observed that the participants made significantly fewer errors in orienting the smartphone correctly when using *SpaceMaze* compared to the *Baseline* application. Additionally, the participants did slightly better or as well as with the *Context-Sensitive* and *Baseline* applications in handling the sensing device, especially avoiding errors like shaking the smartphone, touching or covering the microphone. Also, most of the participants were able to record a measurement with valid length when using *SpaceMaze*. However, there is one significant exception to these positive effects, namely a notable increase in user generated noise. We discuss possible reasons and remedies below.

In order to determine whether the mobile minigame is more enjoyable than other applications for noise measurements and/or offers a better user experience, we interpret both the statistical results as well as the qualitative feedback from our study. The UEQ and SUS measures show that *SpaceMaze* delivers a significantly better user experience than the *Baseline* and *Context-Sensitive* applications regarding stimulation and hedonic quality. For the other dimensions and measures, we noticed varying results, which are in line with our expectations. We expected that our game is more stimulating

and appealing to the users than the *Baseline* and *Context-Sensitive* applications, and at the same time it is understandable that the game loses efficiency and pragmatic quality when implementing the sensified gaming approach. The qualitative analysis gives valuable hints for the future development of the game and reveals potential for further improvement of *SpaceMaze*. From the positive feedback, we can draw the conclusion that our game was enjoyable for most users and can therefore be helpful for motivating users to take an active part in participatory sensing.

6.1. Noise

It is clear from the evaluation results that our approach has shortcomings concerning noise made by a large percentage of users while playing/recording. We believe that different factors contributed to this effect.

First, while we did introduce that the app was a game which was designed to record noise levels during gameplay, we believe that players foremost perceived the application as a game and not as a measurement tool or simply forgot while playing. One the one hand, this is exactly what we intended, as gameplay and real-world task behavior are deliberately aligned to perform the sensing task in the background. On the other hand, this reveals a potential drawback of the Sensified Gaming approach: Since the game is the foreground, peculiarities of the data collection task are naturally not in the user’s working memory. This illustrates that the choice of appropriate game mechanics is key to achieving high data quality. Evidently, in *SpaceMaze*, not all mechanics were chosen optimally.

Second, the noise may partly be due to a novelty effect. Participants were asked to play each the tutorial and the main level exactly once. Since a relevant part of the observed instances of making noise were sounds of frustration or surprise upon death or when struggling with an obstacle, their frequency may be lower if players are a bit more familiar with the game. In this regard, the study situation itself may have contributed additionally in a way, that participants may have behaved differently because an observer was present. We discuss this deeper in the next subsection.

Third, sound is, as an environmental phenomenon, especially difficult to measure with a sensified game. When we look at noise in a broader sense, i.e. not meaning *unwanted sound* but generally *undesired signal fluctuations*, of course noise can occur in any environmental sensing task. In contrast to audible sound, the crowdsourced sensing of other phenomena, such as e.g. UV levels (Turner, Parisi, Igoe, and Amar, 2017) , air quality (Snik, Rietjens, Apituley, Volten, Mijling, Di Noia, Heikamp, Heinsbroek, Hasekamp, Smit, et al., 2014), or weather data (Mass and Madaus, 2014) is unlikely to be problematic in terms of interference with user behavior or game interface channels.

Either way, even for phenomena that are problematic in terms of interference, additional measures can be taken to mitigate the effect. We discuss such an approach below, in subsection *Opportunistic Data Filtering*.

6.2. Observation Bias

We can see that the user generated noise was at least partly caused by the observation situation the participants were in. Often, participants commented on their success or failure in the game or even asked questions, directly addressing the study instructor (e.g. “What does this green thing mean?” or “I think I’m doing something wrong”). The fact that a large number of participants made game-related statements supports

this suspicion. Further investigation is needed to determine if the participants make less noise when playing the game alone or in a situation where they feel unwatched. Table 6 summarizes the statements the participants made. We can see from the statements that the presence of a study instructor may have actually acted as a catalyst: If there were

Table 6. Exclamations and comments given by participants in our evaluation. While participants were not explicitly asked to *Think Aloud*, all of their comments were transcribed by the study instructor.

Swearing and death-related noise	I die again! Oh shit Oh god Damn! That was close ... again! No no no
Comments on the course of the game	I keep touching the walls. I can already see the checkpoint. I am too fast! I've been here before... Almost there!
General observations	This thing rotates. I mustn't touch the red one. The controls are very sensitive. This is the death zone. This 180 degree rotation is... There is a lot of stuff here.
Questions	When will the game continue? Am I supposed to hold it differently? What does this green thing mean? Which way should I go? "You win!" And what happens next?
Other expressions	Laughing Oh oh! Ah Hm...
Justifying personal performance	I am the pro gamer. I'm on a roll now! I am too stupid for that I am doing really bad right now. I think I am doing something wrong I won't finish that in time...

no-one to hear the people vent their frustration, they might not have expressed it out loud in the first place.

6.3. Opportunistic Data Filtering

An important factor when looking at the observed measurement error is that – in contrast to the non-gamified apps – some of the occurred error may actually be irrelevant, as the sensing is performed opportunistically in the background and could be tweaked accordingly. A more sophisticated approach to deal with the noise problem may e.g. be to remove the parts recorded directly after player death from the analysis.

In order to determine whether this would have an effect on the validity of recordings, we re-calculated the error "recording < 30 sec" to incorporate the instances in which players made noise.

In the game log, we wrote a timestamp every time a player died in the game. We observed that the player makes noise, especially out of frustration, shortly before or after he dies. We also recorded a stopwatch log while observing the players, writing a timestamp whenever they made a noise. We assume that, if both the stopwatch log and the game log show a period between two timestamps which is longer than 30 seconds, there is a valid measurement with the required minimum length. This is the case for more than 90 % of the data.

Additionally, most of the errors of the class "microphone directed to user", "microphone covered" and "microphone touched" could also be eliminated. We observed that most errors of the type "microphone directed to user" were caused by one of the two following behaviors. Players either turned the phone directly back after having warped or turned the phone by 360° instead of just 180°, ending up with the microphone pointing towards them again. As a result, the microphone often ended up resting in the palm of the player's hand, touching or covering it. And while the game recognized the wrong orientation and the ship lost its shields, it could be played also in this direction and the level completed without shields. Still, in terms of data collection, it would be easy to discard this data automatically, eliminating the errors caused by wrong orientation.

So overall, by not taking the full middle part of the level to make an ambient audio recording but instead opportunistically cutting out shorter parts that are less likely to be afflicted with measurement error while at the same time conforming to the constraints, the *effective* error rates can likely be strongly reduced and in some cases even eliminated. In order to test this, we calculated the percentage of recordings that were valid (i.e. without noise events and still over 30 seconds long) if we removed all audio within a range of +/- 5 seconds of each "player death" event (see Table 7).

Table 7. When opportunistically selecting the recorded data by editing out a +/- 5 second (respectively 15 second) interval around each "player death" event, the percentage of valid recordings can be substantially increased.

Variant	Participants making noise
Baseline	43 %
SpaceMaze (naive)	90 %
SpaceMaze (opportunistic, editing out +/- 5 sec)	50 %
SpaceMaze (opportunistic, editing out +/- 15 sec)	33 %

While we see a substantial drop in player generated noise, still 50% of all players did make some kind of noise when editing out +/- 5 seconds. However, the percentage is much closer to that observed in the baseline setting. If we increase the time we edit out to +/- 15 seconds, only 33% of recordings remain invalid. This could be because some people continued to comment about the game for a longer time period after crashing their ship. Also, some noises that we attribute to other effects (as discussed above) were expected to remain.

We intend to refine both the game design as well as the evaluation setup in future research to reduce player noise on the one hand and better understand the remaining noise on the other.

6.4. *Design Effort*

An important point worth mentioning is that of design and implementation effort. While we have shown that it is clearly possible to aid the correct execution of sensing tasks through the use of games, designers should balance the effort that goes into the creation of a game with the expected benefits regarding the quality of the collected data. For smaller studies, other approaches like e.g. training may be more sensible than creating a game. On a larger scale, such as the Participatory Sensing scenarios that are the basis for this work, the effort may become less significant in perspective.

6.5. *Ethical Aspects*

Game designer Jane McGonigal stated that “if you use the power of games to give people an opportunity to do something they want to do, then you’re doing good. If you’re using the power of games to get people to do something you want them to do, then you’re doing evil” (Kumar and Herger, 2013). When designing games with a purpose, it would be unethical to coerce or surreptitiously trick the users into measuring without wanting it or even being aware of it. We therefore informed all study participants beforehand that the purpose of the game was to measure ambient noise levels in the background. As it is mandatory in studies and psychological research in general, users need to be able to give their *informed consent*. This is especially true if the task at hand is a location based game, since collected and shared data may additionally raise privacy concerns. Possible ethical issues such as autonomy and deception, unaware participation, and different aspects of privacy and security are discussed at length by Montola, Waern, Kuittinen, and Stenros (2006) in a report on Ethics of Pervasive Gaming in the IPerG project¹, which can also be read as a guideline document for reflecting individual game designs from the ethical point of view. This is something that any sensified gaming system must consider as well.

6.6. *Game-based Approach vs. Education*

Finally, we would like to discuss the benefits and drawbacks of a game-based approach vs. an approach to educate the user on how to use the phone appropriately. A possible area of conflict is the effects that a game approach may have on motivation. It is well-known that adding gamification elements may have negative effects on the engagement of intrinsically motivated users (Mekler et al., 2013). This may especially be the case in a Citizen Science context, which is more than citizens collecting data, but instead is process in which citizens are involved and often drive the whole system. Any educational aspects of such approaches are also not a part of a sensified gaming interface like *SpaceMaze*.

However, we do not propose to present a sensified game as the *only* interface for a measurement system. Instead, it could be one out of different incentive approaches that cater to the needs and interests of different user groups, as suggested by Restuccia et al. (2016). For strongly intrinsically motivated users that do not have an interest in games, a system like *SpaceMaze* may be a problematic choice. A game based design may on the other hand be appealing to a completely different user group and could e.g. help to raise participation levels of generally environmentally conscious young people.

¹<https://cordis.europa.eu/project/rcn/71871/en>

A design that focuses on educating users to correctly perform the sensing task has its benefits and drawbacks as well. While projects that work well in terms of user engagement often make use of intrinsically motivated participants, it makes it harder to reach a large scale and/or coverage. On the other hand, adverse effects on motivation are unlikely in a non-gamified app. Also, related work has shown, that it is not trivial as well to educate users in a scalable and understandable way, so that the variance in measurement behavior is reduced and errors are minimized (Budde et al., 2017a).

We believe that an approach that combines rigorously built “classic” apps that cater to the needs of strongly intrinsically motivated citizen scientist with alternative approaches that facilitate participation of crowdworkers or environmentally interested users with a different social profile would be overall beneficial in terms of data quality, engagement and user experience.

7. Conclusion

In this work, we presented the design and evaluation of *SpaceMaze*, a game intended to guide user behavior in a Participatory Sensing data collection task. As scenario, we selected the recording of ambient noise levels with the internal microphone of a smartphone. Game mechanics were chosen from a set of game design patterns corresponding to the quality constraints of the sensing scenario. In this way, the device is automatically brought into the correct sensing context while the game is being played.

We conducted a field study to examine the behavior of participants when using the minigame *SpaceMaze* for sensing. The objective was to compare the results of *SpaceMaze* to those from the evaluation of a two non-gamified noise level recording apps from previous work. Aside from recording the observed user behavior in terms of errors made, we used standard questionnaires (SUS and UEQ) and qualitative feedback to assess the user experience of the game.

The results of our field study show that *SpaceMaze* can adequately guide the user behavior to create a correct sensing context and reduce human error, such as covering the microphone or shaking the device. On the other hand, we also experienced participants making noise, especially vocalizing game-related remarks, while participating in our study. We propose opportunistically disregarding parts of the recording based on in-game events to mitigate this effect in our discussion. Also, number and type of game-related statements support the suspicion that this effect is in part caused by the observation situation.

In terms of user experience, we notice mixed, but expectable results in comparison to the non-gamified applications. Overall, the game was perceived as more stimulating and scored higher in hedonic quality, while scoring lower in terms of pragmatic quality and efficiency. Still, the qualitative feedback we collected suggests that the minigame is quite enjoyable and offers a good user experience.

For future work, we intend to verify the proposed approach of using additional information to select an opportune time window in which the noise measurement is performed in a separate study with a differently designed observation situation.

Acknowledgements

The authors would like to thank all study participants. This work has been partially funded by the German Federal Ministry of Education and Research (BMBF) as part

of project Software Campus, grant 01IS12051, and the German Federal Ministry of Traffic and Digital Infrastructure (BMVI) as part of project SmartAQnet (Budde, Riedel, Beigl, Schäfer, Emeis, Cyrus, Schnelle-Kreis, Philipp, Ziegler, Grimm, and Gratz, 2017b), grant 19F2003B. Parts of the game design and evaluation have been previously published as part of Matthias Budde’s doctoral dissertation at Karlsruhe Institute of Technology (KIT) (Budde, 2018).

References

- Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. Participatory sensing. *Center for Embedded Network Sensing*, 2006.
- Kyungsik Han, Eric A Graham, Dylan Vassallo, and Deborah Estrin. Enhancing motivation in a mobile participatory sensing project through gaming. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 1443–1448. IEEE, 2011.
- Nicolas Maisonneuve, Matthias Stevens, and Bartek Ochab. Participatory noise pollution monitoring using mobile phones. In *Information Polity*, volume 15, pages 51 – 71. 2010. .
- Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining” gamification”. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15, 2011.
- Elisa D Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N Tuch. Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements. In *Proceedings of the First International Conference on gameful design, research, and applications*, pages 66–73. ACM, 2013.
- Matthias Budde, Marcel Köpke, and Michael Beigl. Robust In-situ Data Reconstruction from Poisson Noise for Low-cost, Mobile, Non-expert Environmental Sensing. In *19th International Symposium on Wearable Computers (ISWC’15)*, 2015. .
- Niels van Berkel, Matthias Budde, Senuri Wijenayake, and Jorge Goncalves. Improving accuracy in mobile human contributions: an overview. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 594–599, 2018.
- Mike Harding, Bran Knowles, Nigel Davies, and Mark Rouncefield. Hci, civic engagement & trust. In *33rd Annual ACM Conference on Human Factors in Computing Systems, CHI ’15*, pages 2833–2842, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6.
- Steven K. Gibb. Volunteers against pollution. *Chemical & Engineering News (C&EN)*, 93 (36), 9 2015.
- Linda See, Alexis Comber, Carl Salk, Steffen Fritz, Marijn van der Velde, Christoph Perger, Christian Schill, Ian McCallum, Florian Kraxner, and Michael Obersteiner. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE*, 8(7), 2013.
- Matthias Budde, Rikard Öxler, Michael Beigl, and Jussi Holopainen. Sensified Gaming — Design Patterns and Game Design Elements for Gameful Environmental Sensing. In *13th International Conference on Advances in Computer Entertainment Technology (ACE2016)*. ACM, 2016. .
- S Andrew Sheppard and Loren Terveen. Quality is a verb: the operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 29–38, 2011.
- Christopher D Wickens, John Lee, Yili D Liu, and Sallie Gordon-Becker. *Introduction to Human Factors Engineering: Pearson New International Edition*. Pearson Higher Ed, 2014. Second Edition.
- Matthias Budde, Andrea Schankin, Julien Hoffmann, Marcel Danz, Till Riedel, and Michael

- Beigl. Participatory Sensing or Participatory Nonsense? — Mitigating the Effect of Human Error on Data Quality in Citizen Science. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(3), 2017a. .
- Francesco Restuccia, Sajal K Das, and Jamie Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *ACM Transactions on Sensor Networks (TOSN)*, 12(2):1–40, 2016.
- Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. *Games with a Purpose (GWAPS)*. John Wiley & Sons, 2015.
- Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.
- Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. Using gamification to inspire new citizen science volunteers. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, pages 18–25. ACM, 2013.
- Kevin Crowston and Nathan R Prestopnik. Motivation and data quality in a citizen science game: A design science evaluation. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 450–459. IEEE, 2013.
- Ioanna Iacovides, Charlene Jennett, Cassandra Cornish-Trestrail, and Anna L Cox. Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. In *CHI’13 Extended Abstracts on Human Factors in Computing Systems*, pages 1101–1106. ACM, 2013.
- Benedikt Morschheuser, Juho Hamari, Jonna Koivisto, and Alexander Maedche. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, 106:26–43, 2017.
- Elena Simperl, Neal Reeves, Chris Phethean, Todd Lynes, and Ramine Tinati. Is virtual citizen science a game? *Trans. Soc. Comput.*, 1(2):6:1–6:39, June 2018. ISSN 2469-7818. . URL <http://doi.acm.org/10.1145/3209960>.
- Nathan Prestopnik, Kevin Crowston, and Jun Wang. Exploring data quality in games with a purpose. *iConference 2014 Proceedings*, 2014.
- Sebastian Matyas, Peter Kiefer, Christoph Schlieder, and Sara Kleyer. Wisdom about the crowd: Assuring geospatial data quality collected in location-based games. In *International Conference on Entertainment Computing (ICEC 2011)*, pages 331–336, 2011.
- Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–21, 2017.
- David R Flatla, Carl Gutwin, Lennart E Nacke, Scott Bateman, and Regan L Mandryk. Calibration games: making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 403–412. ACM, 2011.
- Kathleen Tuite, Noah Snaveley, Dun-yu Hsiao, Nadine Tabing, and Zoran Popovic. Photocity: Training experts at large-scale image acquisition through a competitive game. In *SIGCHI Conference on Human Factors in Computing Systems, CHI ’11*. ACM, 2011. ISBN 978-1-4503-0228-9.
- Donald A. Norman. Design rules based on analyses of human error. *Commun. ACM*, 26(4): 254–258, April 1983. ISSN 0001-0782.
- Fares Kayali, Naemi Luckner, Peter Purgathofer, Katta Spiel, and Geraldine Fitzpatrick. Design considerations towards long-term engagement in games for health. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*, pages 1–8, 2018.
- Staffan Björk and Jussi Holopainen. *Patterns in Game Design*. Charles River Media, Boston, MA, 2004.
- Staffan Björk. Gameplay design pattern collection, 2009. URL <http://www.gameplaydesignpatterns.org/>. Accessed on May 4th, 2016.
- Rajib Kumar Rana, Chun Tung Chou, Salil S Kanhere, Nirupama Bulusu, and Wen Hu. Ear-phone: an end-to-end participatory urban noise mapping system. In *9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 105–116,

- 2010.
- Eiman Kanjo. Noisepay: A real-time mobile phone platform for urban noise monitoring and mapping. *Mobile Networks and Applications*, 15(4):562–574, 2010.
- Immanuel Schweizer, Roman Bärtil, Axel Schulz, Florian Probst, and Max Mühläuser. Noisemap – real-time participatory noise maps. In *Proc. 2nd Intl Workshop on Sensing Applications on Mobile Phones (PhoneSense11)*, 2011.
- Ludovico Antonio Muratori, Paola Salomoni, and Giovanni Pau. Feeling the pack: Strategies for an optimal participatory system to sense and recognize noise pollution. In *2011 IEEE International Conference on Consumer Electronics -Berlin (ICCE-Berlin)*, pages 17–21, 9 2011.
- Matthias Budde. *Distributed, Low-Cost, Non-Expert Fine Dust Sensing with Smartphones*. Doctoral dissertation, Karlsruhe Institute of Technology (KIT), January 2018. URL <https://dx.doi.org/10.5445/IR/1000092736>.
- Staffan Björk, Sus Lundgren, and Jussi Holopainen. Game Design Patterns. In M Copier and J Raessens, editors, *Level Up*, number 3, pages 180–193. Wordware Publishing, 2003. ISBN 1584503548.
- World Medical Association. Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. *Bulletin of the World Health Organization*, 2001, 79, 1964. URL [https://www.who.int/bulletin/archives/79\(4\)373.pdf](https://www.who.int/bulletin/archives/79(4)373.pdf).
- Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer, 2008.
- John Brooke. SUS – A quick and dirty usability scale. *Usability evaluation in industry*, 189, 1996.
- Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, 1988.
- Joanna Turner, Alfio V Parisi, Damien P Igoe, and Abdurazaq Amar. Detection of ultraviolet b radiation with internal smartphone sensors. *Instrumentation Science & Technology*, 45 (6):618–638, 2017.
- Frans Snik, Jeroen HH Rietjens, Arnoud Apituley, Hester Volten, Bas Mijling, Antonio Di Noia, Stephanie Heikamp, Ritse C Heinsbroek, Otto P Hasekamp, J Martijn Smit, et al. Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. *Geophysical Research Letters*, 41(20):7351–7358, 2014. .
- Clifford F Mass and Luke E Madaus. Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bulletin of the American Meteorological Society*, 95(9):1343–1349, 2014.
- Janaki Mythily Kumar and Mario Herger. *Gamification at work: Designing engaging business software*. 2013.
- Markus Montola, Annika Waern, Jussi Kuittinen, and Jaakko Stenros. IPerG – Integrated Project on Pervasive Gaming, Work Package WP5: Design and Evaluation Deliverable D5.5: Ethics of Pervasive Gaming. Public project report, 2006. Release date: October 13 2006.
- Matthias Budde, Till Riedel, Michael Beigl, Klaus Schäfer, Stefan Emeis, Josef Cyrus, Jürgen Schnelle-Kreis, Andreas Philipp, Volker Ziegler, Hans Grimm, and Thomas Gratza. SmartAQnet: Remote and In-Situ Sensing of Urban Air Quality. In *Proc. SPIE 10424, Remote Sensing of Clouds and the Atmosphere XXII, 104240C*, 2017b.