Participatory Sensing or Participatory Nonsense? — Mitigating the Effect of Human Error on Data Quality in Citizen Science

MATTHIAS BUDDE, Karlsruhe Institute of Technology (KIT), TECO, Germany
ANDREA SCHANKIN, Karlsruhe Institute of Technology (KIT), TECO, Germany
JULIEN HOFFMANN, Karlsruhe Institute of Technology (KIT), TECO, Germany
MARCEL DANZ, Karlsruhe Institute of Technology (KIT), TECO, Germany
TILL RIEDEL, Karlsruhe Institute of Technology (KIT), TECO, Germany
MICHAEL BEIGL, Karlsruhe Institute of Technology (KIT), TECO, Germany

Citizen Science with mobile and wearable technology holds the possibility of unprecedented observation systems. Experts and policy makers are torn between enthusiasm and scepticism regarding the value of the resulting data, as their decision making traditionally relies on high-quality instrumentation and trained personnel measuring in a standardized way. In this paper, we (1) present an empirical behavior taxonomy of errors exhibited in non-expert smartphone-based sensing, based on four small exploratory studies, and discuss measures to mitigate their effects. We then present a large summative study (N=535) that compares instructions and technical measures to address these errors, both from the perspective of improvements to error frequency and perceived usability. Our results show that (2) technical measures without explanation notably reduce the perceived usability and (3) technical measures and instructions nicely complement each other: Their combination achieves a significant reduction in observed error rates while not affecting the user experience negatively.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *Ubiquitous and mobile computing design and evaluation methods*; *User centered design*;

Additional Key Words and Phrases: Design Space; Human Error; User Mistakes; Empirical Study; Volunteer Monitoring; Novice Sensing; Crowd Science; Amateur Science; Participatory Sensing; Non-Expert Sensing; User Study

1 INTRODUCTION

Mobile and wearable devices – being always on, always with the user and context-sensitive – present a perfect platform for so-called Participatory Sensing [9]. Projects are highly diverse, ranging from plant observation over data processing (e.g. classifying and labeling data) to sensing environmental phenomena. An extensive survey of Participatory Sensing was presented by Christin et al. [10]. In contrast to the potential of such systems stand the many sources of systematic error that may affect data quality in mobile and wearable Participatory Sensing [6]. That is why, as Bonney et al. point out, "*Despite the wealth of information emerging from citizen science projects, the practice is not universally accepted as a valid method of scientific investigation. [...] At the same time, opportunities to use citizen science to achieve positive outcomes for science and society are going unrealized*" [3]. This goes to the extreme that data from volunteers is considered undesirable by experts or policy makers and may even be

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) as part of *Software Campus* project *FeinPhone* (grant no. 01IS12051), partially funded within the EU FP7 project *Prosperity4All* (grant agreement no. 610510) and partially funded by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) as part of *SmartAQnet* (grant no. 19F2003A).
Authors' address: Karlsuhe Institute of Technology (KIT), Pervasive Computing Systems / TECO, Vincenz-Prießnitz-Straße 1, 76131 Karlsruhe, Germany.

prohibited for official use [15]. On the other hand, research shows that laypersons can collect data of comparable quality to experts, if properly familiarized with the task [38]. The problem is that non-experts are typically not, and thus cannot ensure standardized sensing processes. They may be:

- *Untrained:* unfamiliar with the way the sensing process is intended to be performed,
- *Overwhelmed:* uncomprehending or unable to recall the correct measurement procedure,
- *Inattentive:* not paying attention to all details of the process (esp. likely if participation is extrinsically driven, e.g. through monetary incentives or gamification),
- *Digital immigrants:* not digital natives, i.e. have little or no experience with mobile or wearable technology, or even
- *Malicious:* deliberately trying to influence the measurement process or "to play" the system.

We argue that, in the design of Participatory Sensing tools and applications, the user is still mostly regarded as someone who needs to be motivated and should ideally have a good time using it. Meanwhile it is neglected that in fact often she or he is also an important part of the technical sensing architecture, directly affecting the quality of the generated data or performed task. Harding et al. recently recognized that "*[...] this application domain is poorly understood by most system designers who focus almost exclusively on empowering citizens rather than considering the needs of both citizens and civic authorities and establishing trusted relationships between these stakeholders*" [16]. By designing for both adequate data quality and intelligibility, this trust relationship between users and authorities is strengthened.

This work and its contribution are divided into two parts: In the first part (section 3), we present a series of empirical field observations that we conducted to explore the variance in behavior that non-experts display in different Participatory Sensing settings. We categorize our observations, focusing the perspective of our analysis on the correct execution of the respective sensing process. Subsequently, we gather and categorize mechanisms that can be employed to prevent or mitigate this kind of adverse behavior (section 4). The collected knowledge can be used to guide the development of systems that help non-experts to perform measurement tasks more uniformly and to prevent certain mistakes, thereby increasing data quality.

In the second part (section 5), we present a large field study of an exemplary Participatory Sensing application. Four different designs ('app flavors') are compared to validate the effectiveness of the collected measures, discuss their interplay with user experience and illustrate the importance of making measures understandable to the user.

2 RELATED WORK

This section gives an overview of generally related work. An in-depth discussion of measures that can be employed to increase data quality is discussed in section 4 below.

As already shortly mentioned above, Harding et al. recently recognized that the "*perceived value of civic crowdsourcing applications has remained low*' and that the design space is yet poorly understood [16]. However, their work focuses on engagement and the important trust relationship between different stakeholders, whereas this paper addresses the relationship between non-expert user behavior and data quality. Also centering on motivational aspects, specifically regarding online citizen science platforms, is the work by Yadav et al. [50].

Sensr [20] is one of the rare systems to guide the design of Participatory Sensing tools and deployments. It is a framework for authoring mobile data-collection tools for citizen science. The focus lies on facilitating the process of creating mobile applications for people without technical skills, e.g. through a visual programing environment. As such, it addresses novice programmers more than novice users. Their work also includes a short overview on approaches to ensure data quality. Other than that, individual authors occasionally include discussion on possible non-expert user errors in their work. Klakegg et al. for instance presented a mobile sensing system that enables end-users to perform portable Near Infrared Spectroscopy (NIRS) [23]. As specific challenges, the authors list ensuring the correct distance and angle to the measured object, an even surface, and environmental factors like

avoiding stray light or interference. In our own previous work, we included an overview of different sources of systematic error in mobile sensing [6], one of them being non-expert users. Alagarai Sampath et al. researched how improving the presentation of a task to crowdworkers affects their performance [1], and Dey et al. presented a tool to support building intelligible context-aware systems by exposing the application logic to the user [12]. We included exploiting knowledge about the sensing context as a promising approach to mitigating some non-expert user errors in our discussions below. To the best of our knowledge, no work has yet comprehensively explored the dimensions of non-expert user errors in mobile sensing for citizen science.

Norman very early presented high-level design rules for computer systems based on human error [31]. While his guidelines are still applicable today and also valid for the design of Participatory Sensing, they are very generic. Among other methods, he proposes to "*Use analyses of people's performance in a variety of situations – but especially their errors – to construct an analysis of the appropriate form of human-machine interface that would optimize performance and minimize […] error*". This is what this work explores for Participatory Sensing.

3 PART I: EMPIRICAL DESIGN SPACE EXPLORATION

In the beginning we had little more than the idea that people's sensing behavior is likely to be diverse, considering the fact that the concept of Participatory Sensing emphasizes distributed sensing by everyday users with their personal mobile devices in the public sphere [9], and scenarios generally aim at a large scale. In order to gather information about how much variance people display in their sensing behavior and to what extent "naive" users possess an intuitive knowledge of different sensing tasks, we ran a series of small exploratory field studies that also serve as a baseline for the subsequent research.

Methodologically, we opted against methods like interviews since procedural knowledge is difficult to express verbally [13]. Instead, all studies were run in the field: We conducted three measurement studies and one assembly study, each representative for environmental Participatory Sensing. Subjects were merely put into the context ("*Imagine […] How would you do that?*"), as the natural event of measuring with mobile devices is too rare for true ethnographic observation. No action options or restrictions were specified. The behavior was observed (live and partially additionally in video recordings) by an instructor present throughout the trial run. As participants for the measurement studies, passers-by were approached in urban public spaces (such as a park near a university campus, in the street, etc.). Volunteers were not paid.

To adequately explore the design space and capture the peculiarities of different Participatory Sensing tasks, we selected four different use cases for our empirical field research studies: The first two both required the users to record an audio signal with a smartphone, but differed in the source of the signal. For *Noise Level Monitoring*, participants were asked to take audio measurements with the goal of capturing the outdoor ambient noise level. *Audio Recording and Data Annotation* required users to generate an audio signal themselves, record it and finally annotate the recorded data with a ground-truth label. In the third use case, *Participatory Air Quality Sensing*, participants were asked to use the *iSPEX* camera clip-on module [41] for fine dust measurements. The first two settings both entail tasks that are seemingly simple and that we expected people to have a certain intuition for, even without explicit instructions on "correct" sensing behavior. The third use case was selected to observe behavior in presence of a more complex sensing task and according instructions. A fourth study covered the scenario of *Grassroots Sensing with DIY Hardware*. In this setting, we observed participants while assembling a do-it-yourself (DIY) kit of a sensor station for citizen science air quality monitoring.

The focus of observation in the exploratory studies was the variance in exhibited behavior, 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 [31]: 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. However, we would like to stress

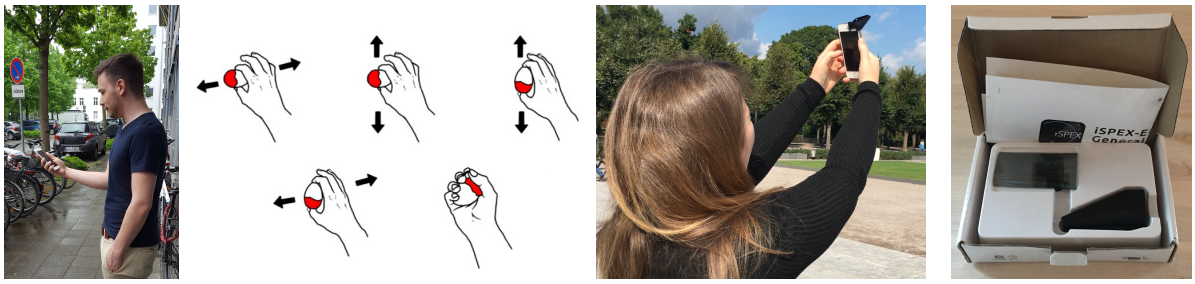
that we do not imply that in these cases the user is to blame. Still, from a technical perspective, they remain errors. An overview of observed human errors is shown and discussed at the end of this section in Table 1, after the presentation of the exploratory studies.

3.1 Exploratory Study 1: Noise Level Monitoring

The underlying scenario for this study is smartphone-based noise pollution sensing. Multiple authors have built phone-based sensing systems for this use case in the past [19, 27, 30, 35, 37]. We selected this application case because it represents the task of measuring an environmental phenomenon with the internal sensors of a standard smartphone. The idea was to get an insight into the varying behavior that we expected to be exhibited for instance by standard users who download a crowdsensing app from the appstore and just intuitively start recording. This use case is the same that also was explored in the final field study in part II of this work.

3.1.1 Participants and Task. Seven participants (four men and three women, ages ranging from 21 to 26 years) were asked to record audio samples representing environmental noise levels using the default *Apple iPhone* audio-recording app (see Figure 1a). All participants were approached in the public sphere of a major city. First, they were given a short introduction into the concept of participatory noise pollution maps. Subsequently, participants were instructed to use the phone to make an audio recording that is representative of the ambient noise pollution level. The only specification on how to do this was the abstract instruction to do it in a way that they felt would yield the highest possible data quality. They were asked to complete a single recording and notify the observing instructor when they thought that they had successfully completed the task. By design, subjects were *not* instructed on correct or incorrect ways to perform the task, in order to not artificially narrow down the range of possible behavior.

3.1.2 Observations. We expected to observe human error in the measurement procedure, and participants' behavior indeed showed both large diversity and scale. As a baseline on what constitutes an error, we adopted the best practices for noise level monitoring from the PDF user guide of the *NoiseTube* project [27]. Of the seven participants, six moved the phone around while recording, causing audible wind-noise in the recording. Six participants held the device too close to their face, breathing into the microphone. Two participants inadvertently covered the microphone with their finger while holding the phone, leading either to muffled recordings or loud scratching noises. In the absence of a timing instruction, the measurement time participants deemed representative of the ambient noise pollution level varied greatly: The duration of the recordings ranged from a few seconds up



(a) (b) (c) (d)

Fig. 1. Exploratory studies 1-3: In study 1 (a), we investigated non-expert user behavior for the use-case of *Noise Level Monitoring* with smartphones. In Study 2 (b), *Audio Recording and Annotation*, we observed a wide range of grip and shake variations. Study 3 (c) involved *Participatory Air Quality Sensing* using the *iSPEX* clip-on module [41] for the *iPhone 5s* (d).

to more than three minutes. Several users actively tracked noise sources, such as cars passing by, and one of them even tried to close in on noise sources by approaching a group of people that were talking and attempted to record the sound of their conversation from a close distance. One participant proceeded to record ambient noise levels without going outside first. Other observed influences were users audibly walking around, scratching themselves loudly and even absurd behavior like talking or whistling (out of boredom) while recording.

3.2 Exploratory Study 2: Audio Recording and Annotation

In the second empirical study, we explored a scenario involving data collection and annotation. This differs from exploratory study 1 in two key aspects: The scenario dealt with additionally performing an activity in contrast to just measuring, as well as sampling an object rather than an environmental phenomenon. This is a more complex setting which also presents more room for human error.

3.2.1 Participants and Task. Thirty-one students (22 men and 9 women, aged between 19 and 32 years) volunteered for this study. We simulated a data collection and labeling task in the following way: Participants were approached on a university campus and asked to shake a *Kinder Surprise* egg and record the produced sound. *Kinder Surprise* is a chocolate egg that contains a small toy inside, which may either be a collectible figure or something that requires assembly. "Experts" claim to be able to determine the type of surprise by the sound of the egg when shaking it. For the use case of building an automated classifier to detect the content of the egg, participants were handed an egg and asked to recording the shaking sound with their smartphone. The specific instruction was to shake the egg and to use their smartphone to record the sound for approximately five seconds and with as little as background noise as possible. Actually opening the egg and assigning a label to the recording was not part of the study.

3.2.2 Observations. While recording, only three participants shook the egg near the microphone, all others shook it somewhere behind or in front of the phone or at one of the sides (left or right). This observation points to an incorrect mental model of sound recording with a smartphone. Participants probably either related the sound recording function to the camera or assumed that the distance between sound source and microphone is irrelevant. Some participants additionally produced unwanted noise by loosely worn watches or bracelets on the arm they used to shake the egg.

Regarding the activity, grip and shaking technique have a large influence and determine the quality of the generated sound. For later classification, it is important that the procedure is ideally performed in the same defined way. The exploratory study shows that, without further instruction, participants hold and shake the egg very differently (see Figure 1b). Most participants held the egg at the long sides and shook it to the peaks (N=13) or perpendicular to the peak axis (N=6). Others hold the egg at the peaks and shook it either to the peaks (N=6) or perpendicular to the peak axis (N=2). Finally, some participants held the egg inside their closed hand (N=4). Half of the participants hold the egg with their dominant hand and the smartphone with the other hand, the other half did it the other way around. Although this is probably not relevant concerning the data quality, it may be of interest in other use cases and generally from the interaction design perspective.

Overall, the exploratory study shows high variance between participants, indicating that the quality of the recorded audio signal can not be guaranteed without more specific instruction.

3.3 Exploratory Study 3: Participatory Air Quality Sensing

The third exploratory study dealt with smartphone-based air quality sensing. Participants were given the *iSPEX* system [41] to measure the fine dust levels in the atmosphere. *iSPEX* is a passive spectropolarimetric add-on for smartphones that uses their camera to determine the levels of atmospheric particles by analyzing the polarization of the light when pointing the sensor add-on at a patch of blue sky.

The measurement process itself is quite intricate. Therefore, the *iSPEX* app includes in-app instructions and mechanisms to guide the user to correct measurements: The principle measurement process is explained both on a one-page paper manual as well as in a tutorial inside the app. Since it is required that the user orients himself so that the sun is in his back, the app calculates the position of the sun for the time of day and the user's location and uses a compass to point him or her in the right direction. The app also tries to detect whether the hardware module is installed and triggers an alert and prohibits measurement if it is not. For a correct measurement, the user slowly needs to raise his or her arm upwards. The app prompts the user in real-time to do this and plays a sound after correct execution to indicate success. All of these features were present in the exploratory study, as we used the standard app available from the AppStore. Because of these measures, sometimes users made errors during the study before the app detected and prevented them and eventually enabled them to correctly follow the measurement procedure. As we are interested in exploring the range of possible behavior, we still included these errors in our recordings and discussion at the end of this section.

We selected this use case because it represents a smartphone-based measurement task for which users are likely to have no intuition at all and that requires additional unfamiliar hardware. Other examples for complex sensing tasks tasks like this are e.g. smartphone-based portable near infrared spectroscopy (NIRS) [23] or light-scattering particle measurements with camera-phones [5].

3.3.1 Participants and Task. In this study, ten pedestrians (seven men and three women, aged 19 to 28, were recruited in a inner-city park close to a European university campus. For the study, they were given an *iPhone 5s* with the most recent app version installed (last updated: Oct, 2015) as well as an *iSPEX* module, complete in box, including the sensor module, an adapter for the *iPhone 5s* and a quick manual (see Figure 1d). The purpose of measuring fine dust using a camera smartphone was shortly introduced to them and they were then asked to use the phone (respectively the app) and the sensor add-on to perform measurements. No specific instructions, e.g. on how to install the add-on were given. All participants owned a smartphone themselves, three of them an *iPhone* and the rest *Android* phones. While the *iSPEX* system only exists for the *iPhone* and thus there was no alternative, our observations do not suggest any impact of providing participants with an unfamiliar phone.

3.3.2 Observations. Six participants immediately hit the "start measurement" button after starting the app, without having installed the *iSPEX* add-on or reading the instructions. Three participants then still left the module in the box at first and tried to take measurements with the app without the hardware add-on being installed, one of them even over and over again for several minutes. Conversely, one user carefully read the included paper manual, installed the module correctly and then started performing the correct procedure, but without having pressed the "start measurement" button first.

All participants encountered problems when attaching the *iSPEX* module to the phone, as they first had to attach the separately packed *iPhone 5s* adapter, and it took a lot of time for them to get it to fit right. Still, for some of them, the app even then wrongly kept displaying the alert "*add-on missing*", preventing them from taking a measurement. One user installed the add-on module facing the wrong way. After receiving an alert, the user corrected this. Another user at first installed the hardware add-on at the bottom side of the iPhone over the microphone, because he read in the manual that he should enable the sound on his phone and therefore thought that the microphone or speaker was used to make the measurements. After reading the in-app tutorial, the user corrected his mistake. The subjects generally had trouble understanding what kind of arm movement was expected from them and that they should continue to raise their arm until the phone was over their head.

Apart from the above errors that were eventually prevented by the mechanisms of the *iSPEX* app, we observed several errors that the app was not able to catch: Of the ten participants, seven tried to take measurements even though no sufficiently large patch of blue sky was visible, five were too close to trees or buildings for a valid measurement. Three participants measured while being seated, without orienting themselves away from the sun.

Overall, a noteworthy observation is that even though the app successfully prevented many types of misbehavior quite reliably by interrupting the measurement attempt, it did not inform the user concerning the reason for the interruption. As a result, three of the ten participants eventually became frustrated and aborted the measurement attempt. As main problem with the use of the app they spontaneously reported the lack of specific feedback (participant #7: "*The app does not tell me what I am doing wrong!*"). Regardless of the source of errors, alerts almost exclusively contained the message "*add-on missing*".

3.4 Exploratory Study 4: Grassroots Sensing with DIY Hardware

The underlying scenario for this study is grassroots environmental monitoring. Around the world, we increasingly witness examples of sensing campaigns that are driven by activists. *Hackspaces* and *Fab Labs*, along with according project descriptions that are widely available over the Internet, have enabled citizens to build and operate sensor stations who could not have done this before. A real-world example for this is the the do-it-yourself (DIY) fine dust sensor by the so-called *OK Lab* of the *Open Knowledge Foundation Germany* in Stuttgart, Germany, a nonprofit organization that advocates open knowledge, open data, transparency, and civil participation. The *OK Lab* provides an online manual[1] that explains the assembly and installment of a sensor station. They also operate and maintain a server to which measurements can be uploaded and an online platform which visualizes the data.

3.4.1 Participants and Task. In this exploratory study, no instructions were given at all. Nine participants (five men and four women, ages ranging from 21 to 57 years) agreed to being observed while assembling the DIY sensing kit of the *OK Lab*. All participants worked at a local newspaper, most of them as journalists, one as technical staff. They chose to build the DIY kit on their own accord. Prior to the observation, they had as a group ordered all parts necessary for the assembly, as listed on the project's website: 10 pcs. each of the WiFi enabled *NodeMCU ESP8266* board, the *SDS011* dust sensor, and the *DHT22* temperature/humidity sensor, as well as some wires, a USB power supply, plastic tubing and a piece of flexible hose. Subsequently, they chose to assemble the individual devices in a group session (see Figure 2a), as explained on the project's website and an FAQ video. Assembling the sensor station required connecting seven wires to the appropriate pins, connecting the flexible

[1]Assembly instructions for the do-it-yourself (DIY) fine dust sensor node: http://luftdaten.info/feinstaubsensor-bauen/.

(a) (b)

Fig. 2. Exploratory Study 4: A group of nine participants (a) each assembled a do-it-yourrself (DIY) sensor station for air quality monitoring (b) in the use case of *Grassroots Sensing with DIY Hardware*.

hose to the air inlet of the sensor, flashing the firmware onto the NodeMCU from a shell, and finally installing the resulting system into the plastic tubing (see Figure 2b).

3.4.2 Observations. When assembling the sensor, each of the participants worked by himself on one sensor kit. Three of the participants at least partially failed to connect the wires to the correct sockets. In four of the assembled sensors, individual wires had slipped out of the sockets. Two of the participants were not able to build the kit by themselves and eventually asked others in the group for help. Three participants lamented the manual being unstructured or even missing steps. To verify that they had successfully assembled the DIY kit, participants connected the completed kits to a power outlet and listened whether the dust sensor's fan started to make a light noise. However, this test did not prevent further error: One participant did not succeed in flashing the firmware without noticing, leaving the stock firmware on the device. Inserting the finalized sensor into the plastic tubing housing went smoothly for all but one participant, who experienced this as being difficult. None of the participants shortened the piece of flexible hose that was connected to the air inlet of the optical dust sensor, which might enable stray light to be reflected into the measurement chamber. When trying to register the sensors in the local WiFi, three of the sensors did not advertise their SSID as described in the instructions and as needed to finalize the configuration. Of the five sensors that were successfully registered, only two included all of the sensor data in their communication. The other three transmitted empty values for either the dust sensor or the temperature sensor, probably due to bad connections or cable breaks.

Since some of the observed errors (and the underlying causes) are not directly evident to the user, after our study, we discussed our findings with an organizer of the *OK Lab*. He reported that they perform regular sanity checks at the back-end, especially on newly registered sensors, as they also observed missing data or that the connectivity of sensors sometimes varies. This apparently happens mostly either because people install them in an inappropriate place outside of the range of their own WiFi or when users do not maintain proper operating conditions, e.g. by switching off their WiFi over night to save energy. To be able to give the users feedback on this, the *OK Lab* has started to transmit the WiFi signal strength along with the sensor data. While the *SDS011* dust sensor comes pre-calibrated and user calibration of the DIY station is not intended in the project, long-term data on stability of the sensor is not yet available and (re-)calibration or sensor replacement may be required [8].

3.5 Analysis of Observed Human Error

The exploratory studies revealed many ways in which participants exhibited human error (both slips and mistakes [31]) in the measurement process, even in seemingly elementary tasks. We collected all our observations in Table 1. The table includes behavior which arguably may not strictly be erroneous per se, but which varied strongly between participants, suggesting a significant effect on the resulting measurement.

Subsequently, we grouped similar instances of behavior into more abstract types of errors in smartphone-based mobile measurements and finally defined six dimensions of human error to form a taxonomy (see Figure 3).

3.5.1 Hardware. This aspect both concerns the employed smartphone (or other personal mobile devices) as well as potentially any other hardware, active or passive, that may be required for the sensing task. In general, one can assume that users will be most comfortable with their own device and that unfamiliar platforms and especially add-ons and external devices are more likely to promote erroneous behavior. This is especially likely if they include intricate assembly and/or maintenance.

3.5.2 Device Handling. This dimension regards the handling of the employed device(s). Requirements may range from virtually non-existing for robust tasks (e.g. recording and submitting textual observations) to tightly constrained procedures that the user is unfamiliar with and/or need to be followed precisely in order to collect meaningful data.

3.5.3 User Activity. The next dimension concerns any behavior of the user that is unrelated to the measurement process and may still affect it, potentially reducing data quality. Mostly, this covers unwanted physical activity and the like. As with device handling, there may be tight constraints regarding this dimension or none at all, depending on the sensing task.

3.5.4 Measurement/Observation. This dimension concerns requirements regarding the recorded observation. Such requirements for high quality data may range from completely free observations to tightly defined constraints, e.g. regarding sample size, annotation requirements, synchronization of different readings etc. The fewer constraints are defined, the more diverse data will probably be collected across participants. Generally, this will likely make comparison and/or data fusion more difficult.

3.5.5 Object/Phenomenon. This dimension covers any requirements regarding the phenomenon or an object that is at the center of the observation. Constraints may range from basically none to looking at precisely defined aspect of a specific object. The boundaries between this dimension and the handling of the device may overlap, e.g. when a certain alignment between device and object is required.

Table 1. Observed non-expert behavior in our exploratory studies (study 1: *Noise Level Monitoring*, study 2: *Audio Recording and Annotation*, study 3: *Participatory Air Quality Sensing*, study 4: *Grassroots Sensing with DIY Hardware*).

Dimension	Type	Example Behavior	
Hardware	Faulty installation	clip-on module missing	(see study 3)
		clip-on module incorrectly installed	(see study 3)
	Incomplete / wrong assembly	incorrectly built DIY kit	(see study 4)
	Undocumented / surplus parts	polarization foil	(see study 3)
		incomplete manual	(see study 4)
	Missing maintenance	operating conditions not ensured	(see study 4)
	No / faulty calibration	un- or decalibrated sensors	(see study 4)
	Faulty device association	sensing device not paired, loss of data	(see study 4)
Device handling	Wrong orientation	user turns around own axis while measuring	(see study 1)
		microphone pointing in the wrong direction	(see study 1)
		moving device in wrong angle	(see study 3)
	Wrong height or distance	microphone too close to face	(see study 1)
		arm not extended	(see study 1)
	Unwanted device movement	user noisily moves device around	(see study 1)
		user shakes device instead of (or along with) egg	(see study 2)
	Covering sensor	user has finger on microphone	(see study 1)
User Activity	Generating noise	talking, whistling	(see study 1)
		coughing, clearing one's throat, breathing noisily	(see study 1)
		scratching, clapping	(see study 1)
		unwanted noise (watches rattling)	(see study 2)
	Unwanted user movement	walking around	(see study 1)
	Phone use	making a call or texting	(see study 1)
Measurement/Observation	Wrong sample properties	recording too short	(see study 1)
	Wrong amount of samples	only one measurement instead of two	(see study 3)
	Not actually sensing / recording	user forgot to press recording button	(see study 3)
	False or no annotation	wrong label assigned by mistake	(see study 2)
		no position data, as GPS disabled / no signal	(see study 1)
Object/Phenomenon	Wrong object handling	wrong grip or shake	(see study 2)
	Wrong object/phenomenon	no blue skies visible	(see study 3)
	Wrong alignment	shaking egg nowhere near microphone	(see study 2)
		user follows noise source and tries to get very close	(see study 1)
Environment/Context	Inappropriate weather	noisy wind in audio recording	(see study 1)
		no blue skies visible	(see study 3)
	Indoors instead of outdoors (or v.v.)	attempting to measure ambient noise levels indoors	(see study 1)
	Environmental disturbances	stray light entering sensor	(see study 4)
	Wrong time and/or place	measuring at wrong location or point in time	(see study 4)

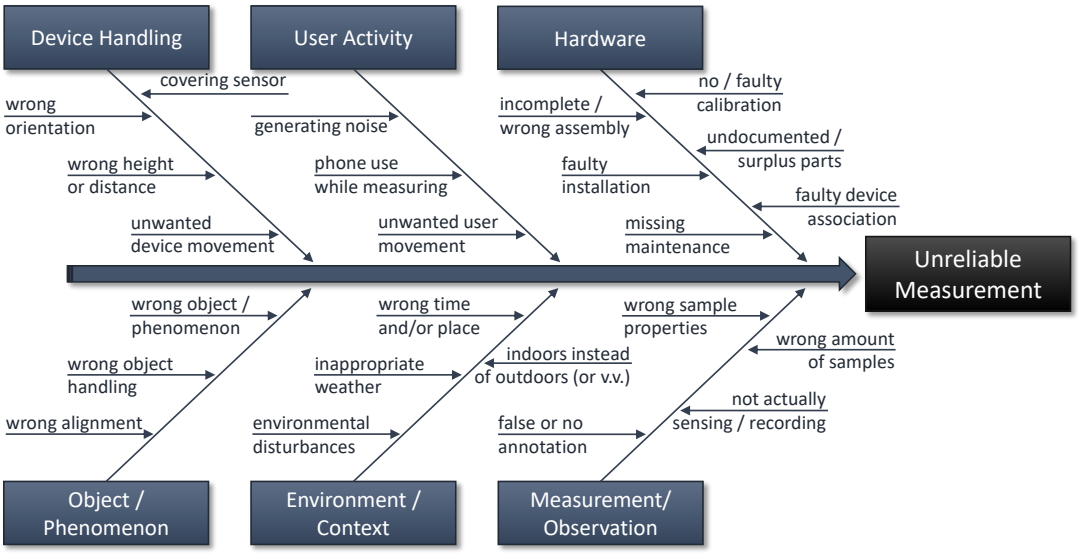


Fig. 3. Ishikawa diagram of the identified dimensions through which users may affect the quality of the measurement result.

3.5.6 Environment/Context. The last dimension concerns the environmental context[2] of the user. Measurements may be robust to external factors and be allowed anywhere and anytime or again, tightly constrained and well-defined.

4 ENHANCING DATA QUALITY

After having analyzed and compiled ways in which participants' behavior may adversely affect data quality in Participatory Sensing, we look at ways to prevent the undesired behavior or mitigate its effects in this section. Research on mobile sensing recognizes the need to ensure viable readings from low-cost sensors [24], even though the focus is seldomly placed on the effects that are caused by non-expert users. Many general approaches exist in the literature, some of which are applicable to typical Participatory Sensing scenarios and some of which are not. Whether or not a technique is suitable or not depends on a variety of aspects, such as the specifics of the task at hand, the scale of the deployment, etc. This section discusses classes of possible countermeasures gathered from literature review. Table 2 summarizes the results.

Participant Selection is an approach that has been used in different fields to identify and separate suitable personnel from such that is unfit for a task. Wickens et al. review different methods of identifying people who are likely to perform successfully along with different measures of ability, albeit with a focus on assigning people to jobs [47]. However, by definition, Participatory Sensing addresses everyday users, which may make pre-selection an undesirable step. Additionally, at larger scale, screening may become prohibitively expensive.

The most intuitive approach to ensure that users perform a task correctly is *training* [26, 38, 39]. Thelen et al. reported that "*numerous studies have demonstrated that volunteers can successfully perform basic data collection tasks when given a half day or more of practical field training.*" [43]. This highlights the biggest drawback of training sessions: A lot of resources (experts, facilities, etc.) are needed and the approach does not scale. Slightly

[2]We are aware that in Ubiquitous Computing, the term *Context* by itself usually includes aspects that are already covered in other dimensions here, e.g. activity. Context has i.a. been defined as "*any information that can be used to characterize the situation of an entity*" [11]. In contrast, environmental context here is meant more narrow.

different forms of training that do not require the user to keep a mental model of the process are *instructions* (e.g. manuals or tutorials). The key difference to training is that instructions are typically given in writing or otherwise fixed form (video, etc.) which is used to make the non-experts to understand the measurement process. Understanding is defined as the ability to hold and process all elements that define the measurement process simultaneously in working memory [42]. However, working memory is extremely limited in capacity [29] and in duration [34], in particular for novel information that needs to be processed in a novel way [42]. That is, people might fail to understand or completely recall new material if it is sufficiently complex, as may be the case in Participatory Sensing. As we have shown in our exploratory studies, even a seemingly easy task like recording an audio signal involves a complex measurement process for the user. Also, pure manuals are of little help, as people tend to not read them [33], especially if they do not encounter problems, as would be the case in a badly but successfully performed measurement process. On the other hand, instructions can be given much "closer" to the actual task (spatially and temporally). In shorter form and *in-situ*, instructions provide an advantageous approach, up to providing a step-by-step walkthrough.

Another popular approach are *reputation* systems [18]. There are different flavors, ranging from picking users based on their reputation or skill level [36] (cmp. *selection* avove) over assessing it beforehand [44] to building it through data analytics. However, this again requires some kind of ground truth determined by expert users

Table 2. Overview of possible measures to improve the data quality in mobile non-expert sensing.

Measure	**Advantages**	**Disadvantages**
Participant Selection	domain expertise / prior knowledge	usually requires thorough analysis, high ressource cost; selection success hard to verify
Training	best / covers everything; trainer can assess success	high resource cost, experts needed continuously
Instruction (manual)	clear	mentally demanding; passive access
Instruction (in-situ)	very clear; temporally close	requires some sort of display
Reputation	good for sorting out single users; helps against malicious intent	can de-motivate users; may be hard to build
Verification	ensures data quality	infeasible for many tasks; after-the-fact;
Expert Reviews	ensures data quality	infeasible for many tasks; experts needed continuously; after-the-fact
Redundancy	very simple	only robust against statistical error, not systematic; may be difficult to achieve; bad readings still contribute
Outlier/Anomaly Detection	eliminates implausible readings	prevents capturing "true" anomalies
Bayes Filter	adapts to available data	needs basic models and multiple readings
Repetition	simple	only robust against statistical error, not systematic; needs to be triggered somehow
Context Recognition	potentially fine grained control over measurements	only certain class of errors; maybe technically difficult
Reconstruction	extremely robust	only applicable in special cases
Data Design	provides structure;	mostly for textual data;
Feedback	supports user in verifying correct procedure himself; almost always possible	may overwhelm or frustrate user if not carefully balanced
Gamification	motivates; increases hedonic quality; may enhance measurement frequency	may distract from sensing task: can demotivate intrinsically motivated participants

or a series of campaigns, making it an intricate option. In Participatory Sensing systems, individual readings can often not be re-evaluated and the classification of them as being correct or wrong after-the-fact is often infeasible, making reputation levels difficult to build. Additionally, "ranking users can backfire" [22], influence the participants' motivation, and paradoxically lead to the best performing participants quitting, as they would feel they had "won the game".

Verification of data entries is another approach to increase data quality. Gardliner et al. [14] differentiate between verified and direct citizen science. Entry verification can either be approached automatically by using some sort of computational recognition or simple sanity tests. The advantage of this sort of verification is, that it can be performed in real-time and the user can be prompted before leaving the area, as proposed by Burke et al. ("*Did you really just see 40 diesel trucks go by in five minutes?*") [9]. In community-based data validation [48], instead of revisiting their own data, participants verify data from their peers. Another form of verification are *expert reviews* [20] of data. They have the disadvantage that data has already been collected and can only be discarded, as the analysis takes place after-the-fact.

Computational approaches are diverse. The simplest ones are of a statistical nature: *redundancy* [20] and/or *repetition* [38] both lead to multiple instances of the same data which can then in turn be processed, e.g. to remove outliers. These approaches only work, if the underlying assumption holds that the overall error is non-systematic, i.e. people will on average perform the task correctly. However, as we have seen in our exploratory studies, there are certain errors which the majority of people tend to make. More sophisticated approaches like *outlier/anomaly detection* or *Bayes filtering* take the structure of the data into account. The drawbacks of filtering out anomalies is that the smoothing makes the approach less suitable for highly dynamic phenomena. If only few data points are available or no model can be constructed, filtering is also not applicable.

A different way of computationally addressing procedural errors is *context recognition*. Mechanisms may be as simple as detecting whether the GPS receiver is turned on or the acceleration sensors of a device pick up movement when there should be none, to integrating full-fledged activity recognition. A robust way to deal with different types of error afflicted data is signal *reconstruction* from noise [6]. However, it is not generally applicable, as the measured phenomenon must be modelable as particles, among other constraints.

An interesting approach is *data design*, i.e. using HCI methods not only to design interfaces, but also to assess the needs of data consumers to collect reliable, standardized and overall more useful data [21]. However, this works for observations that are reported in a free form (e.g. textual), but not so much for pure sensor data. Additionally, models have been developed to exchange, revise and merge structured offline data, e.g. from contributions that are accomplished via paper [40].

Finally, one of the most universal mechanisms is *feedback*. Since we assume that the user actually is interested in collecting and submitting high-quality data[3], it is important to make the measurement process as transparent as possible. Feedback (e.g. on the correct execution of a step, etc.) can greatly contribute to this understanding.

Some of the discussed approaches can be combined with *Gamification* techniques. In this way, the location-based game *GeoSnake* [28] has been used to boost verification rates, the game *PhotoCity* gamifies training [45] and it has been proposed to use game contexts to ensure correct execution of a sensing task [7].

5 PART II: FIELD STUDY

The previous observation studies revealed a surprising diversity in user behavior, which is likely to yield a high variance in data quality. Thus, it is important not only to unify the measurement process per se but also to guide the user behavior, in particular in Participatory Sensing. As already discussed in section 4, there are two main strategies to achieve this goal. First, users could be trained or instructed to show the required behavior or,

[3]We disregard malicious users here, as we are convinced that someone determined to willingly submitting false data will find a way to do so.

second, the correct user behavior could be supported by technical measures. At the same time, user experience is important in order to keep the user motivated to participate in citizen science.

We argue that implementing purely technical measures to prevent certain errors may increase data quality but at the same time may have adverse effects on the general user experience. Conversely, focusing purely sensing on ease of use and an understandable process may still result in users keeping to make certain kinds of errors. We hypothesized that the combination of both will lead to substantially better systems. To test this hypothesis, four functional variants (flavors) of a mobile sound recording app were developed. The four app flavors mainly differed in (a) the way participants were instructed how to record an audio signal properly and (b) the technical measures supporting correct recording (see below).

5.1 Participants

A total of 535 passing-by pedestrians were recruited to volunteer in the study (opportunity sample; 209 women, 321 men; 5 missing values, mean age: 30 years, age range: 18-76 years). The experiment was conducted face-to-face in field (e.g. on the street or in parks). The overall education level of the participants was rather high. About one third had a university degree. About half of the participants ($N = 267$) had a technical or natural scientific background, the others had a commercial ($N = 80$), medical ($N = 24$), juristic ($N = 7$), pedagogic ($N = 53$), administrative ($N = 15$), manual ($N = 22$) or artistic ($N = 18$) education. Fourty-nine participants worked in other branches ($N = 41$) or did not indicate their line of business ($N = 8$).

All participants were well familiar with using mobile phones, most of them (96.6%) also using a smartphone. Almost all participants had used their smartphone to call someone (94.2%) or had sent messages (93.6%) before. Most participants had also in the past sent emails (76.6%) or taken pictures (81.7%). More important in the context of the study, only fewer than half of the participants had used their smartphone to record videos (47.5%), speech (26.0%) or music (16.4%). None of the participants were experts in Participatory Sensing, only 18 of them (3.3%) had already participated in a previous Participatory Sensing study. Overall, the sample was heterogeneous with regard to age and education and representative for app users.

All participants gave their written informed consent and did not receive any compensation for their participation. The study was carried out in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

5.2 Material, Study Design, and Procedure

5.2.1 Procedure. At the beginning of the experiment, participants were introduced to the scenario. They were told to imagine that they were part of a community that tries to build a noise map of their city. For that purpose, they would use an app on a smartphone to record the ambient noise level of their environment with the goal to achieve data quality as high as possible. Participants were then asked to record ambient noise levels with a quality as good as possible. They were given one of the recording apps, which they could use as long as they wanted and undertake so many trials until they felt that they recorded a good signal. No further instruction about the measurement process or the handling of the app was given. After completing the task, they filled in the User Experience Questionnaire (UEQ) and the System Usability Scale (SUS) and gave qualitative feedback to evaluate the app. In addition, they answered some questions about their smartphone usage behavior as well as a few questions concerning demographics and their habits regarding technology use. The whole experiment took about 15-20 minutes per participant. An overview of the session structure is shown in Figure 4.

5.2.2 App Flavors. Four functional variants (flavors) of a mobile sound recording app were developed (see screenshots in Figure 5). The four app flavors mainly differed in (a) the way participants were instructed how to record an audio signal properly and (b) the technical measures supporting correct recording (see Figure 4). App 1 (*Basic*) was a simple one-button app that only allowed starting and stopping a recording. In app 2 (*Basic+*) a short tutorial at the beginning provided detailed instructions of how to avoid erroneous behavior while recording an

audio signal and how to use the app. The presented best practices included, for example, instructions to point the microphone of their smartphone away from their body and avoid shaking it. App 3 (*Premium*) helped avoiding common errors by providing feedback such as status indicators, e.g. when no GPS signal was available or error messages when the user shook their smartphone too strongly (see Figure 5). By flipping the orientation of the display upside-down, the user was forced to rotate the smartphone so that the microphone pointed away from the body. In app 4 (*Premium+*) the same short tutorial was included as in app 2. Aside from the sound sample, all app flavors automatically recorded certain events (tutorial usage, recording times, etc.) in a logfile for evaluation.

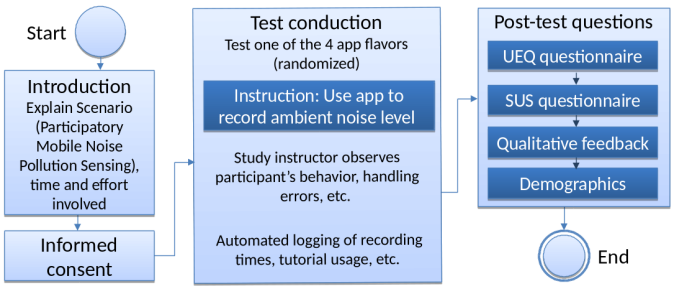
5.2.3 Study Design. The study consisted of four experimental conditions, i.e. the four app flavors. Test conditions were evenly assigned to participants in a between-subject design, i.e. each participant used only one of the four app flavors (Basic: $N = 123$; Basic+: $N = 137$; Premium: $N = 130$; Premium+: $N = 145$).

5.2.4 Collected Data. We captured (i) the number and types of errors that users made with different variants of the citizen science app, as well as (ii) the user experience while performing the sensing task. User errors were recorded in categories by the study instructor in an observation protocol, following the previously identified error dimensions (cmp. Figure 3 above). Erroneous behavior was defined by the same best practices as in our exploratory studies. User experience and usability were measured with the German versions of the User Experience Questionnaire (UEQ) [25] and the System Usability Scale (SUS) [4]. The SUS is a standardized questionnaire with ten short questions that primarily measures the usability aspect of a product. The UEQ is supposed to measure user experience in a wider scope and consists of 26 bipolar items that are assigned to the six scales *Attractivenes*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*. In addition, qualitative feedback was collected.

5.3 Data analysis

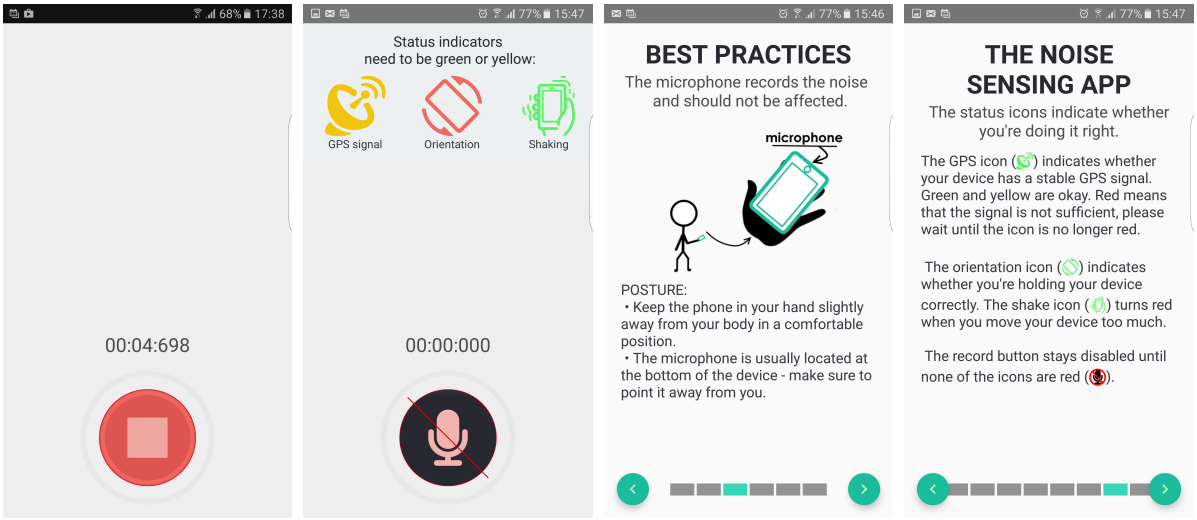
Data was analyzed by comparing only those app flavors with each other which were of interest regarding our study goals: (i) To analyze the effect of instruction (i.e. in-app tutorial) alone, *Basic* and *Basic+* were compared; (ii) the effect of technical measures alone (without explanation), was evaluated by comparing *Basic* and *Premium*; (iii) to explore the difference between the effects of instruction and technical measures, *Basic+* (instruction only) and *Premium* (technical measures only) were compared; and (iv) *Basic* and *Premium+* were compared to show the complementary effect of instruction and technical measures.

Statistically, four independent t tests were computed, separately for usability and user behavior data. Because of multiple comparisons, p values were adjusted according to Bonferroni. For this correction, the alpha level



	Basic	*Basic+*	*Premium*	*Premium+*
Animated Button	×	×	×	×
Time Display	×	×	×	×
Button Vibration	×	×	×	×
In-App-Tutorial		×		×
In-App-Manual		×		×
Interface upside-down			×	×
Check: GPS on?			×	×
Check: Orientation?			×	×
Check: No shaking?			×	×
Check: Rec. time?			×	×
Alerts			×	×

Fig. 4. Session structure of the field study (left) and the four different experimental conditions, i.e. app flavors (right): A simple 1-button recording app (*Basic*) serves as baseline for our study. The *Premium* flavor features multiple measures to improve data quality, including sensor-based verification of parts of the sensing context. *Basic+* and *Premium+* additionally display an in-app tutorial.



(a) *Basic* / *Basic+* (b) *Premium* / *Premium+* (c) Tutorial: best practices (d) Tutorial: app manual

Fig. 5. Screenshots: The simple *Basic* flavors only feature an animated recording button and a time display (a). In *Premium* flavors, recording is disabled unless certain constraints are met (b). Additionally, the *Basic+* and *Premium+* versions each feature an in-app-tutorial explaining best sensing practices (c) and the app (d).

(usually $p < .05$) is divided by the total number of pairwise comparisons. Then each of the p values are compared to that shrunken value of alpha. To improve the readability of the text, we did not reduce the level of significance but report Bonferroni adjusted p values that allow a direct comparison to a level of significance of $p < .05$. This is common practice, for example, by statistic software such as SPSS. SPSS multiplies each of the actual p values by the total number of possible pairs. That is, a test can be considered statistically significant if the reported $p < .05$. The descriptive results are presented in Figure 6 (user behavior during the measurement process) and Table 3 (usability and user experience).

Qualitative data was analyzed with a content analysis. We started with small clusters of semantically related comments which were further grouped on the basis on the categories of the DIN EN ISO 9241/110 (software ergonomics). Some clusters were related to very specific aspects of the app (e.g., the tutorial). On the most abstract level, we grouped the clusters based on pragmatic quality (perspicuity, efficiency, and dependability), hedonic quality (novelty and stimulation), and attractiveness in order to compare them to the results measured with the UEQ. It is important to note that the apps were developed with the objective to observe the measurement process, i.e. the functionality of the app was limited[4]. This was also noticed by the participants. While these comments were omitted from analyses, they do explain the low ratings for all four app flavors in some of the categories of the UEQ.

5.4 Study Results

The study results are reported sorted by (i) the effect of instruction alone (*Basic* vs. *Basic+*); (ii) the effect of technical measures alone (*Basic* vs. *Premium*); (iii) the difference between the effects of instruction and technical

[4]Specifically, after sensing, there was no visualization of the measurement result in a map, as dB value, or the like. This was by design, as we focused on measurement error and did not want to solicit qualitative feedback on the visualization in this study.

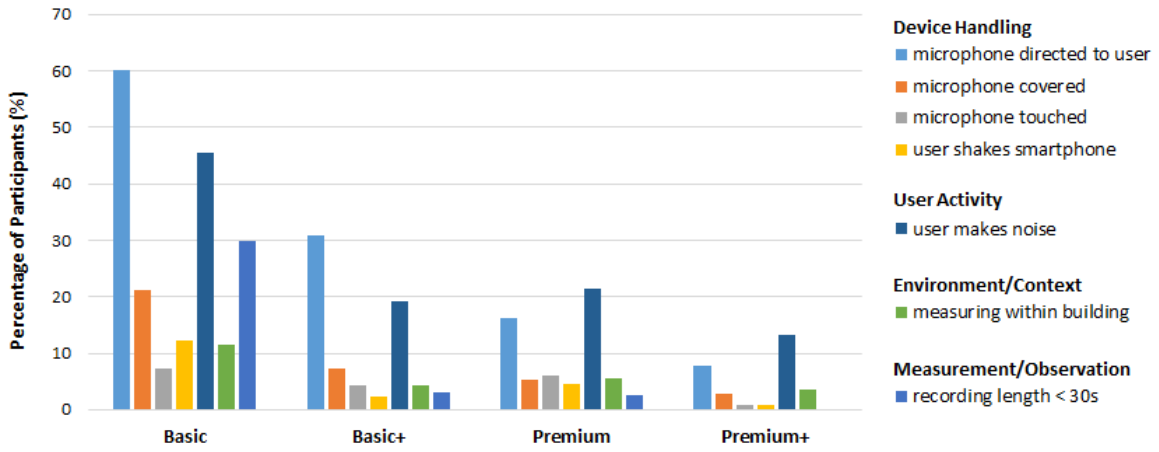
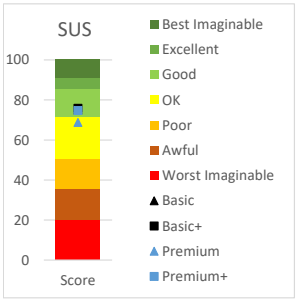


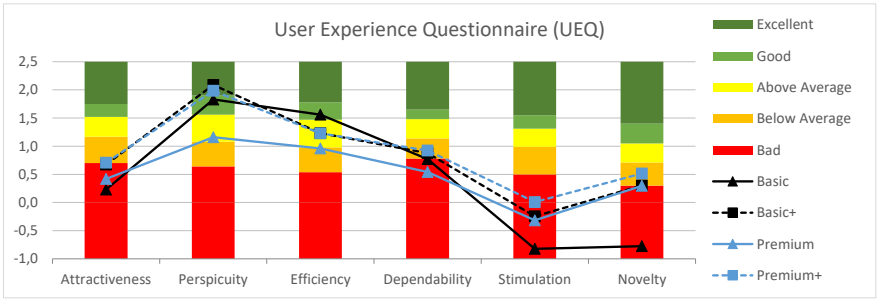
Fig. 6. Measurement errors observed in the four different test conditions (i.e. app flavors).

measures (*Basic+* (instruction alone) vs. *Premium* (technical measures alone)); and (iv) the complementary effect of instruction and technical measures (*Basic* vs. *Premium+*). A summary is given at the end of this section.

5.4.1 Effect of Instruction Only. Adding a tutorial (*Basic+*) to a simple one-button app (*Basic*) improved the measurement process in terms of errors made. In particular, the microphone was less frequently directed to the user, $t(258) = 4.845$, $p < .001$, or being covered, $t(258) = 3.280$, $p = .008$; users shook the smartphone less often, $t(258) = 3.224$, $p = .008$; the required recording duration was achieved more frequently, $t(243) = 6.126$, $p < .001$;

Table 3. Results of the SUS and UEQ questionnaires (Att = attractivity; Per = perspicuity; Eff = efficiency; Dep = dependability; Sti = stimulation; Nov = novelty). While technical measures alone (*Premium*) reduce user errors in a similar way as instructions alone (*Basic+*, see Figure 6), the perceived usability is lower.





	SUS	Att	Per	Eff	Dep	Sti	Nov
Basic	76	0.2	1.8	1.5	0.8	-0.9	-0.9
Basic+	76	0.7	2.1	1.3	0.9	-0.1	0.3
Premium	**69**	0.4	**1.2**	1.0	0.5	-0.2	0.3
Premium+	75	0.7	2.0	1.2	1.0	0.1	0.6

and the user produced less interfering noise, $t(258) = 4.781$, $p < .001$. However, despite the general reduction in errors, users still made some errors.

Adding a tutorial (*Basic+*) that described the measurement process did not change the usability of the app. The SUS score remained the same, $t(246) = 0.066$, $p = 1.000$, as well as perspicuity, $t(258) = 2.225$, $p = .112$, efficiency, $t(258) = 1.840$, $p = .268$, and dependability, $t(258) = 1.259$, $p = .836$, as measurement of pragmatic aspects of user experience. The perceived hedonic quality increased, i.e. stimulation, $t(258) = 4.764$, $p < .001$, and novelty, $t(258) = 7.466$, $p < .001$. Overall, an app with a tutorial was perceived as being more attractive than without it, $t(258) = 3.713$, $p < .001$.

The qualitative data allowed a more detailed understanding of these results. With regard to the pragmatic quality, most users mentioned that both one-button apps were easy and fast to use and it was easy to understand how to use it. The observed change in hedonic quality was also reflected in the qualitative data. While 14 users mentioned that using the basic app was boring or uninteresting, this number reduced to 5 when adding a tutorial. More important, 15 participants even rated the app as being innovative or that they felt fun while recording the audio signal. The low value in attractiveness was probably due to the overall look of the app. Although most participants liked the clear and simple design, it also appeared empty to them.

5.4.2 Effect of Technical Measures Only. The technical implementations were also able to improve user behavior: The microphone was less frequently directed to the user, $t(250) = 8.227$; $p < .001$, or covered, $t(250) = 3.785$, $p < .001$. The recording duration was less frequently below 30 sec, $t(234) = 6.143$, $p < .001$; and the user produced less interfering noise, $t(250) = 4.110$, $p < .001$. In this case, the usability decreased: The SUS score was significantly lower, $t(242) = 3.022$, $p = .012$, and also two of the UEQ scales that are supposed to measure the perceived pragmatic quality of a product showed reduced values, i.e. perspicuity, $t(250) = 4.151$, $p < .001$, and efficiency, $t(250) = 4.526$, $p < .001$; dependability did not change, $t(250) = 1.855$, $p = .264$.

Table 4. Statistical study results. The *Change* value represents the observed difference between the two conditions, whereby negative values represent an aggravation relative to the *Basic* app, whereas positive values represent improvements. The change is statistically significant if $p < .05$. Because of multiple comparisons, p values were adjusted according to Bonferroni.

		Basic* vs. *Basic+		***Basic* vs. *Premium***		***Basic* vs. *Premium+***		***Basic+* vs. *Premium***	
	Variable	Change	p value	Change	p value	Change	p value	Change	p value
User Experience	SUS	0.1	1.0	-6.4	.012	0.6	1.0	6.5	.004
	UEQ Attractiveness	0.5	<.001	0.3	.192	0.6	<.001	-0.3	.208
	UEQ Perspicuity	0.3	.112	-0.6	<.001	0.1	1.0	-0.9	<.001
	UEQ Efficiency	-0.2	.268	-0.6	<.001	-0.3	.064	-0.3	.020
	UEQ Dependability	0.2	.836	0.2	.264	0.2	.348	-0.4	.004
	UEQ Stimulation	0.8	<.001	0.6	<.001	0.9	<.001	-0.1	1.0
	UEQ Novelty	1.2	<.001	1.2	<.001	1.4	<.001	0.0	1.0
User Behavior	Microphone to user	28.8	<.001	44.7	<.001	52.5	<.001	15.9	.008
	Microphone covered	13.8	.008	15.7	<.001	18.3	<.001	1.9	1.0
	Microphone touched	2.9	1.0	1.1	1.0	6.6	.032	-1.8	1.0
	Shaking	10.0	.008	7.5	.128	11.5	<.001	-2.5	1.0
	Length < 30 sec	24.1	<.001	23.0	<.001	27.1	<.001	-1.0	1.0
	Inside building	7.0	.156	5.9	.376	7.9	.068	-1.1	1.0
	User makes noise	26.6	<.001	23.8	<.001	32.2	<.001	-2.7	1.0

The qualitative data show that participants liked the responsive design. They mentioned that the technical measures helped them avoiding measurement errors. However, without an explanation, not all participants did understand how to use the technical measures correctly. Although participants were able to complete the task, the task itself (e.g., the goal or the action steps) often remained unclear.

5.4.3 Difference Between Instruction and Technical Measures Only. The behavior when using technical measures to guide the measurement process was comparable to that observed when a tutorial was provided, all $t(264) < 1.0$, $p = 1.0$, except for the direction of the microphone. Here, the technical implementation worked more effectively, $t(264) = 3.088$, $p = .008$. However, the usability was much better when providing a tutorial, all $t(264) > 2.829$, $p < .021$, while the perceived hedonic quality was similar, all $t(264) < 1$, $p = 1.0$.

The qualitative data show that the tutorial enhanced the comprehension of the task (i.e. the measurement process per se), whereas the technical measures mainly help the users to handle the smartphone correctly.

5.4.4 Complementary Effect of Technical Measures and Instruction. The previous comparisons showed that the technical implementation and the instruction led to a similar user behavior, but the usability was reduced dramatically when technical features were implemented without explaining them to the user. The users made fewer errors in all categories, all $t(265) > 2.690$, $p < .033$, except for the location, $t(265) = 2.417$, $p = .068$. This might be a statistical artifact of the study, as about 80% of the participants were already outside of a building when asked to participate in the study. Because we recorded the initial location and the location of the actual measurement, we were able to assess how many participants changed their location. When using the *Basic* app, 10 of 24 participants (41.7%) left the building, whereas 24 of 29 participants (82.8%) left the building when using the *Premium+* app. This difference was statistically significant, $\chi^2(1, N = 53) = 9.642$, $p = .002$.

When combining the positive effects of the technical measures with those of the instruction, compared to the *Basic* app, the usability did not significantly decrease, measured with the SUS, $t(252) < 1.0$, $p = 1.0$, and the three scales of the UEQ that are supposed to measure the pragmatic aspects of a product, i.e. perspicuity, $t(265) < 1.0$, $p = 1.0$, efficiency, $t(265) = 2.417$, $p = .064$, and dependability, $t(265) = 1.724$, $p = .348$. However, the hedonic aspects, i.e. novelty, $t(265) = 9.704$, $p < .001$, and stimulation, $t(265) = 6.000$, $p < .001$, increased.

In the qualitative data two clusters were observed that grouped together likes and dislikes of (i) the the technical support of the user (which was described as responsive behavior of the app), and (ii) the tutorial of the *Premium+* app. Overall, the responsive behavior was perceived as very positive. None of the users commented not to understand the function of the technical support. The *Basic* and the *Premium+* app were described as easy and fast to use. However, only when a tutorial was provided, the participants reported to comprehend the task. They even reported that the felt that they cannot make any mistakes. However, it should be noted that the hedonic quality remained low and the task itself was still experienced as being boring.

5.4.5 Summary of the results. The study showed that (i) technical measures as well as instructions reduced observed error rates, (ii) technical measures without explanation reduced the perceived usability and user experience, and (iii) technical measures and instructions nicely complement each other. The instruction enhanced the comprehension of the task (i.e. the measurement process per se), whereas the technical measures mainly help the users to handle the smartphone correctly.

6 DISCUSSION AND LESSONS

In this section we discuss our previously presented findings and summarize takeaway messages.

6.1 Empirical Taxonomy

In the course of empirically building our taxonomy of human error in Participatory Sensing and Citizen Science (as summarized above in Figure 3) we had many discussions concerning the comprehensiveness of the collection, its value to others as well as the selection of the use cases for building it.

We of course are aware that there are many areas of mobile sensing and current devices' capabilities afford for the collection of multiple types of data, some more prone to user error than others. In order to build an abstract characterization of problems, we attempted to select specific but representative use cases that entail different aspects of Participatory Sensing. On an abstract level, many applications fit the topics of the four selected use cases (sensing phenomena, handling objects, annotating data, using additional unfamiliar hardware/devices and assembling equipment). As such, we are convinced that the taxonomy will be a useful ressource for designers and researchers concerned with improving non-expert data collection from mobile devices.

Regarding the comprehensiveness of the taxonomy, we are not aware of analytical work on errors in Participatory Sensing that could have served as a baseline for comparison with our empirical approach. We have however recently become aware of a characterization of all factors that make up uncertainty in measurement processes in industrial testing, authored by the *German Association of the Automotive Industry (VDA)* [46]. They distinguish influences between pertaining to the measurement system (measurement standard, mounting fixture, measuring equipment and measurement parameters) or the measurement process (environment, object, methods, and operator). Our own taxonomy can be mapped well to most of these aspects, indicating a high degree of completeness. The most notable difference is that in industrial processes, the human operator has no power over the process or the sytem, while in mobile sensing, the user directly or indirectly influences all of its aspects.

6.2 Balancing Data Quality and Usability

The results of our study show that technical measures alone can already help to significantly reduce human error in Participatory Sensing. However, an interesting finding is that built-in automated mechanisms for improving data quality may be detrimental to the user experience. While technical measures alone and instructions alone each seem to have a roughly similar effect in terms of error reduction, technical measures without explanation notably reduce the perceived usability of the app, potentially frustrating the users.

6.2.1 Instructions vs. technical measures . Taken at face value, this seems to suggest that citizen science mobile sensing app designers may be better off focusing on making clear instructions (and embedding them into the app) before designing mechanisms to nudge user behavior towards accurate data collection. Looking closer, it is not that simple. While quantitavely, both technical measures and instructions reduce error by a similar amount, they address different kinds of adverse behavior. Obviously, some errors are not preventable using technical measures: not making noise in sound sensing is a good example. While e.g. algorithms for voice activity detection have existed for a long time [17], it is nigh impossible to automatically decide if the detected speech is part of the ambient noise or an artifact of improper measurement procedure. On the other hand, we can simply instruct participants to not talk while recording (and hope that they do). Another difficulty in using technical measures was illustrated in our empirical study with the *iSPEX* system (see Exploratory Study 3 above). Mechanisms that automatically verify certain aspects of the sensing context may be designed poorly or too restrictive. When the user is certain he has done everything correctly and the system insists that this is not the case, this results in frustration. Approaches to improve this include either imposing less strict constraints or making the app more intellgible, e.g. by making it possible for the user to better comprehend and maybe even override automatic decisions, thus providing control over the context-aware application to the user [2].

Instructions on the other hand have their drawbacks as well. Designing adequate instructions is a complex task [49], and on top of that, people tend to not read manuals [33] and/or skip tutorials. A downside of written instructions also is readibility, particularly outdoors: one of the participant in Exploratory Study 1 performed

their measurements in direct sunlight, having a lot of trouble reading the screen and moving around a lot as a result. Varying levels of literacy may also need to be considered when desiging tutorials in certain countries. Solutions to these problems may include using different kinds of instructions (icons, videos) or displays (audio instructions, vibration, etc.).

6.2.2 Remaining Uncertainty. A finding that one may overlook in the face of the observed improvements is that even in the best of our four cases, a significant amount of participants still exhibit erroneous behavior: More than 12% of the participants made noise while recording despite being instructed not to and roughly 8% of the users still attempted to sense with the interface of the app being upside-down and after being explicitly prompted to flip their phone. This suggests that even if considerable effort is placed into reducing human error, a certain amount of afflicted data with questionable quality may always have to be expected. Of course, this can not simply be generalized. An important aspect in that is to not only think about types of errors and the frequency of their occurrence, but also rate them regarding their severity in terms on their effect on the quality of the measurement. The qualitative feedback that participants gave in our field study also provided some insights into why some errors still might have remained. The requirement to measure over a longer time period, e.g. more than 30 seconds as in the current study, was perceived as being boring and uninteresting. This phase may have contributed to the occurrence of users making noise. Here, some strategy to reduce boredom, such as embedding the sensing task in a game context [7], might work to mitigate this type of erroneous behavior.

6.2.3 Recurring Users / Long-term Behavior. In our study, we sampled every participant exactly once, so in that sense, we have shown that the implemented measures can effectively mitigate errors caused by people who have not performed the task before. This is important as it can significantly lower the threshold for newcomers and infrequent participants. But what about recurring participants? Without further measures, it can be both argued that performing a task repeatedly may either increase data quality or not. On the one hand, revisiting data collection activities should reduce slips, i.e. errors in carrying out the intention [31], over time. Mistakes (errors in the intention) on the other hand are more likely to be repeated.

Concerning the effect of instructions and technical measures, we would argue that technical measures are likely to maintain their positive effect over time and maybe even combat slips that would have otherwise occurred, e.g. due to decreasing motivation. Regarding instructions, it is likely that in the long run, people will read the tutorial less carefully or not at all anymore. However, the instructions do not only have a teaching character but also serve as a reminder. Not revisiting them may result in an increasing amount of mistakes in the long run, as people may forget individual steps or mix them up. This also relates to certain design choices that may be important for the long-term experience: Should tutorials be skippable? If so, would people skip them already the first time without reading them? If not, will people be annoyed by them in the long run? One approach could be to reshape the tutorial over time to slowly transform from a tutorial to a shorter reminder.

6.3 Stakeholders

Both for grassroots movements, like the campaign of the *OK Lab* described in Exploratory Study 4, and "top-down" approaches, often driven by experts, it is important to incorporate the interests of all relevant parties. If e.g. activists gather information concerning environmental stress, it is crucial to talk to civic authorities early on, because in the end they will judge whether they accept the data quality as being adequate and make decisions concerning data use. Conversely, organizers need to keep the interests of participants in mind, as people do not want to be instrumentalized and reduced to data collection tools.

However, there is not only a range of stakeholders (participants, organizers, researchers, authorities,...), in our discussions so far we have also seen different positions within these groups. For example, we have heard of municipalities that are strongly interested in the possibilities that distributed low-cost sensing may offer and that

actively work towards integrating such approaches into their current monitoring networks. On the other hand, there are civic authorities that are either wary, or even actively work against crowdsensing projects, possibly out of fear that the gained information may result in financial burden. But even among participants, we have seen both citizens that are e.g. interested in improving air quality and those that are opposed, because they feel that the pollution does not really affect them, but measures to combat it probably would, like bans on motorized traffic. Things get even more complicated if extrinsic motivation to participate is present, like gamification or monetary incentives, because the primary goal of the participant may not be data collection anymore.

In the end, we believe that project coordinators need to work closely with volunteers and system designers, and ultimately successful systems will have to involve a creative collaboration between the variety of actors and stakeholders, which also addresses non-technical aspects like social, cultural, and political issues [32]. Human-centered design, starting with observations on how participants err in real-world situations can be used to build software systems for high-quality Participatory Sensing in an incrementalist approach, that in turn can help to establish mutual trust among stakeholders.

6.4 Takeaway Lessons

Gardliner et al. [14] argue, that while the risk of low data accuracy is present in citizen science, the cost-effectiveness of crowd-sourced science compared to the conventional approach outweighs the risk, if properly handled. We summarize our key findings regarding the proper handling of risks from our studies and discussions, respectively the recommendations for building Participatory Sensing systems in the following list:

- Be aware of the diversity of human error and the effect it may have on data quality.
- Analyze people's errors in order to adopt user interfaces that help to prevent them.
- Address different classes of error with appropriate measures (e.g. instructions and/or technology).
- Design for intelligibility: It's not just what technology can do, but also how a user perceives it.
- Do not overreach, too strict constraints will frustrate the user.
- Involve stakeholders early on to balance required data quality and necessary complexity.

While we encourage designers to shift the perspective towards the correct execution of the measurement process, data quality should not be the only goal. Rather, the focus on data quality should complement proven user-centered design processes. Our study shows that technical measures and instructions nicely complement each other. Our findings highlight the criticality of balancing technological features and their perceived ease of use, a fact which both practitioners and researchers need to be aware of.

Given the problems inherent in accurately predicting system performance in real-world environments, conducting small exploratory studies has proven to be an easy way to collect erroneous behavior specific to the task at hand. The error dimensions presented in this work should provide a useful starting point for system designers developing interfaces and interaction in a way that minimizes the occurrence of human error and thus leads to more uniform and overall better data quality in Participatory Sensing. This is the theme of this paper.

7 CONCLUSION

This paper focuses on the interplay between non-expert user behavior and data quality in Participatory Sensing. To foster a deeper understanding of citizen science tools, it explores the design space of mobile citizen science sensing tools and applications, with the focus on human error. We have presented an empirical taxonomy of errors exhibited in non-expert smartphone-based sensing, based on four small exploratory studies. A large field study that compares instructions and technical measures to address these errors shows that technical measures without explanation notably reduce the perceived usability and the combination of technology and instructions achieves a significant reduction in observed error rates while not affecting the user experience negatively.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) as part of *Software Campus* (grant no. 01IS12051), partially within the EU FP7 project *Prosperity4All* (grant no. 610510) and partially by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) as part of *SmartAQnet* (grant no. 19F2003A). We would like to thank all study participants, investigators and experts for their participation, Volker Ziegler for providing us with an *iSPEX* module, as well as all reviewers for their valuable comments.

REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. In *32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 3665–3674.

[2] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[3] Rick Bonney, Jennifer L Shirk, Tina B Phillips, Andrea Wiggins, Heidi L Ballard, Abraham J Miller-Rushing, Julia K Parrish, et al. 2014. Next steps for citizen science. *Science* 343, 6178 (2014).

[4] John Brooke. 1996. SUS – A quick and dirty usability scale. *Usability evaluation in industry* 189 (1996).

[5] Matthias Budde, Pierre Barbera, Rayan El Masri, Till Riedel, and Michael Beigl. 2013. Retrofitting Smartphones to be Used as Particulate Matter Dosimeters. In *International Symposium on Wearable Computers (ISWC'13)*.

[6] Matthias Budde, Marcel Köpke, and Michael Beigl. 2015. Robust In-situ Data Reconstruction from Poisson Noise for Low-cost, Mobile, Non-expert Environmental Sensing. In *International Symposium on Wearable Computers (ISWC'15)*. ACM, 179–182.

[7] Matthias Budde, Rikard Öxler, Michael Beigl, and Jussi Holopainen. 2016. Sensified Gaming – Design Patterns and Game Design Elements for Gameful Environmental Sensing. In *13th Int. Conference on Advances in Computer Entertainment Technology (ACE2016)*.

[8] Matthias Budde, Lin Zhang, and Michael Beigl. 2014. Distributed, Low-cost Particulate Matter Sensing: Scenarios, Challenges, Approaches. In *ProScience*, Vol. 1. (Proc. 1st Int. Conf. Atmospheric Dust (DUST2014)).

[9] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. 2006. Participatory sensing. *Center for Embedded Network Sensing* (2006).

[10] Delphine Christin, Andreas Reinhardt, Salil S. Kanhere, and Matthias Hollick. 2011. A survey on privacy in mobile participatory sensing applications. *Journal of Systems and Software* 84, 11 (2011). Mobile Applications: Status and Trends.

[11] Anind K. Dey. 2001. Understanding and Using Context. *Personal and Ubiquitous Computing* 5, 1 (2001), 4–7.

[12] Anind K. Dey and Alan Newberger. 2009. Support for Context-aware Intelligibility and Control. In *CHI '09*. ACM, 859–868.

[13] J St BT Evans. 1988. The knowledge elicitation problem: a psychological perspective. *Behaviour & Information Technology* 7, 2 (1988).

[14] Mary M Gardiner, Leslie L Allee, Peter MJ Brown, John E Losey, Helen E Roy, and Rebecca Rice Smyth. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Frontiers in Ecology and the Environment* 10, 9 (2012), 471–476.

[15] Steven K. Gibb. 2015. Volunteers Against Pollution. *Chemical & Engineering News (C&EN)* 93, 36 (Sept. 2015).

[16] Mike Harding, Bran Knowles, Nigel Davies, and Mark Rouncefield. 2015. HCI, Civic Engagement & Trust. In *33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2833–2842.

[17] John D Hoyt and Harry Wechsler. 1994. Detection of human speech in structured noise. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, Vol. 2. IEEE, II–237.

[18] Kuan Lun Huang, Salil S Kanhere, and Wen Hu. 2010. Are you contributing trustworthy data?: the case for a reputation system in participatory sensing. In *Modeling, analysis, and simulation of wireless and mobile systems*.

[19] Eiman Kanjo. 2010. Noisespy: A real-time mobile phone platform for urban noise monitoring and mapping. *Mobile Networks and Applications* 15, 4 (2010), 562–574.

[20] Sunyoung Kim, Jennifer Mankoff, and Eric Paulos. 2013. Sensr: Evaluating a Flexible Framework for Authoring Mobile Data-collection Tools for Citizen Science. In *Computer Supported Cooperative Work (CSCW '13)*. ACM, 10.

[21] Sunyoung Kim, Christine Robson, Thomas Zimmerman, Jeffrey Pierce, and Eben M. Haber. 2011. Creek Watch: Pairing Usefulness and Usability for Successful Citizen Science. In *SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 2125–2134.

[22] Eli Kintisch. 2011. How to Grow Your Own Army of Citizen Scientists. (24 Feb. 2011).

[23] Simon Klakegg, Chu Luo, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2016. Instrumenting Smartphones with Portable NIRS. In *Adj. Proceedings UbiComp'16, Workshop in Ubiquitous Mobile Instrumentation (UbiMI)*.

[24] N.D. Lane, E. Miluzzo, Hong Lu, D. Peebles, T. Choudhury, and A.T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (2010).

[25] B. Laugwitz, T. Held, and M. Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.

[26] Kurt Luther, Scott Counts, Kristin B. Stecher, Aaron Hoff, and Paul Johns. 2009. Pathfinder: An Online Collaboration Environment for Citizen Scientists. In *Human Factors in Computing Systems (CHI '09)*. ACM, 239–248.
[27] Nicolas Maisonneuve, Matthias Stevens, and Bartek Ochab. 2010. Participatory noise pollution monitoring using mobile phones. In *Information Polity*. Vol. 15. 51 – 71. Issue 1.
[28] Sebastian Matyas, Peter Kiefer, Christoph Schlieder, and Sara Kleyer. 2011. Wisdom about the Crowd: Assuring Geospatial Data Quality Collected in Location-Based Games. In *International Conference on Entertainment Computing (ICEC 2011)*. 331–336.
[29] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 91 (1956), 81–97.
[30] L. A. Muratori, P. Salomoni, and G. Pau. 2011. 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)*. 17–21.
[31] Donald A. Norman. 1983. Design Rules Based on Analyses of Human Error. *Commun. ACM* 26, 4 (April 1983), 254–258.
[32] Donald A. Norman and Pieter Jan Stappers. 2015. DesignX: Complex Sociotechnical Systems. *She Ji* 1 (2015), 83–106. Issue 2.
[33] DG Novick and K Ward. 2006. Why Don't People Read the Manual?. In *Int. Conference on Design of Communication (SIGDOC '06)*. ACM.
[34] L. Peterson and M. Peterson. 1959. Short-term retention of individual verbal items. *Journal of Experimental Psychology* 58 (1959).
[35] Rajib Kumar Rana, Chun Tung Chou, Salil S Kanhere, Nirupama Bulusu, and Wen Hu. 2010. Ear-phone: an end-to-end participatory urban noise mapping system. In *9th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 105–116.
[36] Sasank Reddy, Deborah Estrin, and Mani Srivastava. 2010. Recruitment Framework for Participatory Sensing Data Collections. In *Pervasive Computing*. LNCS, Vol. 6030.
[37] I. Schweizer, R. Bärtl, A. Schulz, F. Probst, and M. Mühläuser. 2011. NoiseMap – real-time participatory noise maps. In *PhoneSense'11*.
[38] L. See, A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner. 2013. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS ONE* 8, 7 (2013).
[39] S Andrew Sheppard and Loren Terveen. 2011. 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*. 29–38.
[40] S. Andrew Sheppard, Andrea Wiggins, and Loren Terveen. 2014. Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data. In *17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, 1234–1245.
[41] F Snik, JHH Rietjens, A Apituley, H Volten, B Mijling, A Di Noia, S Heikamp, RC Heinsbroek, OP Hasekamp, and JM Smit. 2014. Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. *Geophysical Research Letters* 41, 20 (2014).
[42] John Sweller. 2002. Visualisation and instructional design. In *International Workshop on Dynamic Visualizations and Learning*.
[43] Brett Amy Thelen and Rachel K. Thiet. 2008. Cultivating connection: Incorporating meaningful citizen science into Cape Cod National Seashore's estuarine research and monitoring programs. *Park Science* 25, 1 (2008).
[44] A. Truskinger, Haofan Yang, J. Wimmer, Jinglan Zhang, I. Williamson, and P. Roe. 2011. Large Scale Participatory Acoustic Sensor Data Analysis: Tools and Reputation Models to Enhance Effectiveness. In *E-Science*.
[45] Kathleen Tuite, Noah Snavely, Dun-yu Hsiao, Nadine Tabing, and Zoran Popovic. 2011. PhotoCity: Training Experts at Large-scale Image Acquisition Through a Competitive Game. In *SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 10.
[46] VDA-QMC. 2011. Messsystem und Messprozess sind zweierlei. *QZ* 56, 5 (2011).
[47] Christopher D Wickens, John Lee, Yili D Liu, and Sallie Gordon-Becker. 2014. *Introduction to Human Factors Engineering: Pearson New International Edition*. Pearson Higher Ed. Second Edition.
[48] Andrea Wiggins and Yurong He. 2016. Community-based Data Validation Practices in Citizen Science. In *19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, 1548–1559.
[49] Patricia Wright. 1981. "The instructions clearly state…" can't people read? *Applied Ergonomics* 12 (September 1981), 131–141. Issue 3.
[50] Poonam Yadav and John Darlington. 2016. Design Guidelines for the User-Centred Collaborative Citizen Science Platforms. *arXiv preprint arXiv:1605.00910* (2016).

Received February 2017; revised May 2017; accepted June 2017