Lessons from Failures in Designing and Conducting Experimental Studies – a Brief Anecdotal Tutorial

Matthias Budde
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
budde@teco.edu

Micheal Beigl
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
michael@teco.edu

Anja Exler
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
exler@teco.edu

Andrea Schankin
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
schankin@teco.edu

Till Riedel
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
riedel@teco.edu

UbiComp/ISWC'17 Adjunct , September 11–15, 2017, Maui, HI, USA
ACM 978-1-4503-5190-4/17/09.
https://doi.org/10.1145/3123024.3124392

Abstract
User studies are an important part in the design of socio-technical systems. However, Computer Science students and practitioners in Ubicomp often have not been trained in designing and conducting experimental studies. To help bridging this gap, this paper presents a (very) short tutorial on how (not) to design and conduct experimental studies and analyze the gathered data. This is done through examples of failure from case studies, which illustrate pitfalls, challenges, and some approaches to solving them in an anecdotal fashion. Most of the failures were committed by ourselves or students at our lab, some anecdotes were added from external references or personal conversation.

Author Keywords
Design; User Studies; Failures; Error; Lessons; Ubicomp

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces – Evaluation/methodology

Introduction
In the design of today's complex socio-technical systems, a proposed strategy is *muddling-through*, i.e. iteratively designing and testing isolated parts in order to converge towards a working system [16]. This naturally involves conducting a lot of observations, user studies, etc. The bar for

User Studies in a Nutshell

Hypothesis: Before doing anything, you should be clear on what you are looking for.

Population: Select and recruit your participants appropriately for your research question.

Study Design: Identify dependent and independent variables, test conditions. Design the study procedure in a way that avoids confounding factors.

Test Run: Pre-test your study with one or two guinea pigs to identify deficits and (if applicable) make sure all technology works stable.

Conduction: Plan time slots, collect informed consent, do not vary procedure (environment, instructions, etc.) unless that is part of the study design.

Analysis: Reflect on the power your results have (or don't have), make sure that the conclusions you draw do not overgeneralize and discuss the external validity of your findings.

that has never been so low, but the devil is in the details. In Computer Science – while people are usually well-educated in probabilistic theory and statistics – we surprisingly often find a lack of understanding user studies. We thus decided to compile this collection of failures in designing, conducting and analyzing experimental studies, mostly intended for students and researchers with a technical background and little experience in human-centered design. Proper study design and execution fills entire text books [14, 21], so by no means do we intend to give a comprehensive overview. We rather present important points on experimental studies (see box *User Studies in a Nutshell*) along with some vivid examples (highlighted in grey) to help them "sink in".

First Things First [Hypothesis]

Before running any study, we should have some kind of an idea in our heads on what we would like to find out. It pays to explicitly ask (and answer) oneself the question "*What am I looking for?*". Typically, it is one of two things:

1. an answer to a question, i.e. to *test a hypothesis* or
2. exploration, i.e. to *generate hypotheses* (to test later)

In order to be scientifically sound, we cannot decide this after-the-fact. While it is valid to conduct exploratory research without having a clear hypothesis, we obviously can not use the same experiment to generate and prove the same hypothesis. Otherwise, correlations that the data analysis shows could be spurious. The crucial factor in this is that instead of testing for a specific correlation, we test for *any* correlation – drastically increasing the probability of a random hit. In principle, testing multiple hypotheses within one experiment is completely valid, as long as we factor this in when analyzing the data (see below), since testing multiple hypotheses increases the probability that at least one of them yields a false positive result.

A nice example that illustrates the above points is an intentionally badly designed study in which researchers "proved" that eating a chocolate bar a day will speed up weight loss by 10% [2]. While the results of the study actually showed accelerated weight loss, only 15 people were studied in three different groups and as many as 18 different factors were measured. By this, the researchers intentionally increased the likelihood of a "statistically significant" result – without caring whether the final headline would read "Chocolate helps you lose weight" or "Chocolate lets you sleep better".

Who You're Gonna Call? [Population]

Once we have decided what we are looking for, we must select an appropriate sample from a population. As we saw from the previous example, this is one of the things in life where size does matter[1]. How many participants you need depends on many different factors, however there are some general rules of thumb:

- Exploratory studies intended to discover problems, e.g. for usability testing of a novel design, can already yield a high data confidence with ten users [8].
- The smaller you estimate the *effect size* of what you wish to observe, the more participants you need.
- A *between-subjects* design generally requires more subjects than a *within-subjects* study.

Aside from the size of the sample, its composition (age, gender, background, experience, etc.) is usually of importance. A quick analysis of all user studies presented at

[1] There are also tools to help calculate the necessary sample sizes like e.g. *G*Power* [7] (http://www.gpower.hhu.de/)

Some Important Terms in Experiments

Independent Variables:
All variables in a process that are manipulated.

Dependent Variables:
Variables that we measure to see the effect of the changes in the independent variable.

Confounding Variables:
Variables that influence both a dependent and an independent variable, causing a meaningless correlation between them.

Treatment:
All levels of an independent variable.

Within-Subjects Design:
The same group of participants serves in more than one treatment (i.e. repeated measures).

Between-Subjects Design:
Different groups of participants serve in each treatment.

MobileHCI 2011, Henze found a bias towards male participants in their twenties with a technical background, which he attributes to researchers going for the *sample of convenience*, i.e. relying on "students and guys from the lab" as participants [9]. This need not be a problem per se, but you should keep possible bias in mind. Getting a nice sample for a study often costs a lot of resources. Engaging your creativity and thinking out-of-the-box can produce innovative ways to recruit participants. If you are interested in a certain characteristic in your study group, it might pay to think where your research interest and interests of your target group overlap.

> In a project on navigation in Zürich, Switzerland, scientists were interested in sampling a large number of international participants from all around the world. In order to quickly and easily recruit them, the researchers advertised their study at the local hostel and offered to cover the rent for one night in exchange for the participation in their study.

> A different example of a win-win situation is a study which aimed at collecting inertial sensor data on eating behavior [19]. As compensation for their participation in the study, researchers "paid" the participants a free lunch, the only condition being that sensor data was recorded while they ate it.

An approach to boost numbers that is typically used in psychology education is that students have to participate in studies as part of their curriculum. Another option is using micro-task markets like *Amazon mTurk* to recruit large numbers of subjects through micro payments [12]. However, both of the above approaches involve strong extrinsic incentives for participation, which is not without its risks.

> In an *mTurk* study we ran, we tasked participants to work on a data cleaning task using a specially designed interactive web application [4]. In the background we collected a lot of data on click events and the like to later assess the system design and participant performance. When analyzing the data after the study, we found that ~5% of the people had tried to cheat us. While the task was "work for 45 minutes to finish as much as you can", one log file e.g. showed two minutes of activity, then 41 minutes of break, followed by another two minutes of work.

> In a recent large user study on errors people make when using smartphones for environmental sensing [5], we altered the psychology approach described above a bit. Instead of using students as participants, they were (as part of their curriculum) tasked with collecting data from four participant (for four experimental conditions) each. The study design was thoroughly prepared and documented and about 200 student instructors were trained beforehand. What we did not anticipate is that some of them attempted to hand in faked results by filling in questionnaires themselves, because they were only interested in obtaining the credit points. Again, these attempts were filtered out relatively easily, based on application log data and checking the internal validity scores of the employed standard questionnaires [13].

There's No Data Like More Data? [Variables]

While in the two previous anecdotes, the collection of additional, possibly redundant data eventually saved the study, it generally depends whether it pays to collect more data or not. The intended analysis usually sets a lower bound on which and how much data is required (algorithms, sta-

Significance

Null Hypothesis: This is the default position that the relationship we are testing for does not exist. Rejecting or disproving the null hypothesis is what we try to do with experimental studies.

Statistical Significance: Generally, a statistic is significant if we are reasonably sure that the relationship the statistic shows is true, i.e. does not just show due to chance (if the null hypothesis were actually true). Statistical significance does not necessarily mean practical significance. A finding may be true without being important. That is subject to interpretation.

Significance Level: What "being reasonably sure" means is determined before collecting the data. We set a probability threshold below which we assume the null hypothesis to be false. This is the significance level, which is typically set to 5%.

tistical significance, etc.). Beyond that, it depends on how "expensive" additional data is. Automatic event logs can be considered to be relatively cheap, as long as the entries are cleanly formatted and documented. Adding an additional sensor may offer interesting additional options to explore.

> We revisit the example of the study where participants were given a free meal in exchange for data from a wrist-worn sensor collected while eating [19]. If memory serves, the authors regretted in retrospect not having equipped each participant with a second wrist-worn sensor band, because they later came up with the question if the activity of eating could also have been detected from the resting hand.

On the other hand, collecting additional sensor data can have unwanted implications for deployment, collection, storage, transmission, synchronization or compatibility. Generally, it increases complexity both for study designers and participants. Asking yourself what burden you place on your subjects is important, even more so for survey data. If you come to the conclusion that there is no harm in collecting some additional data, e.g. for exploration, you should collect the least important data towards the end of a session. In that way, possible effects of decreasing motivation on the participants' side will not affect your core results.

When talking about more data, an important distinction is (again) that of additional data vs. additional variables. If by adding a data source, we in fact add a *dependent variable* to our study design, we need to correct for this in the analysis. Otherwise, we are well on our way to practicing *p-hacking*, like we saw in the first example on multiple hypotheses. Each additional dependent variable increases the probability that we find some correlation by chance (see above).

I've Got a Plan. [Study Design]

A lot of the complexity of a study both on the user and the designer side can be dealt with by proper planning. One thing that is frequently overlooked is knowing what expectations participants of a study might have themselves and what can reasonably be "imposed" on the subjects. Sometimes it may make more sense to conduct multiple study sessions for different aspects. This may of course introduce other problems: Participants may not be available twice, environmental conditions change, learning effects kick in, instrumentation may differ between the sessions and the overall organizational effort has to be invested a second time. A compromise is to balance time and reliability, e.g. by reducing observations to a lower but acceptable level.

> In an EEG study [18], we were convinced to having worked out the perfect study plan. Unfortunately while the plan was indeed well-suited for collecting exactly the data we needed, we failed to determine the time budget for a session. In the end, participants would have needed to invest six hours, which was infeasible. As a compromise, we reduced the number of trials to cut the time budget in half.

The selection of technology in the planning phase should also always happen bearing the needs and capabilities of participant groups in mind.

> Study designers selected a well-designed and powerful web-based survey tool. Unfortunately, only English was available as language, which the user group was unlikely to understand. Since nobody had thought about the user group, no resources were available in the project for translation. Ironically, the topic of the project was accessibility.

An important part of planning also involves ethics approval, privacy considerations and getting the participants' informed consent. Aside from following ethical guidelines, there may be important rules set by your organization or even government concerning proper procedure.

> We heard from a study in the US for which – after having received ethics board (IRB) approval – researchers corrected a typo in the consent form. As a result, they had to throw away all collected data, as the legal situation was that *any* alteration after-the-fact invalidated IRB approval.

Computer Says No. [Test Run]

Testing the study design before going forward is always advisable. Whenever technology is involved, it becomes especially critical to think about the differences of the test setup and the real-life study environment that is to follow.

> For a smartphone study we used the *Google Places API*, which was configured to allow up to 1000 requests per day. While this was enough for our tests, it unfortunately was not for the study that followed.

> Another study app of ours [1] used both the microphone and speaker at the same time, which worked fine for the devices we tested with. In the final study, however, users owned previously unseen devices that did not support it and crashed as a result.

> In a third study, the employed ESM app was not configured to start automatically after a device restart. Many participants restarted their phones at some point and no more data was collected from them.

But even if the collection itself runs smoothly, obtaining the data in the end needs to be thought through as well.

> In a study app, the collected data was stored internally and obtained by a manual data export after the study had finished. However, some participants had uninstalled the app before debriefing, deleting the data with it.

Finally, it is important that you do the test run well before the final study is scheduled. Otherwise you may not have time to fix possible issues or worse: introduce new ones.

> One day before a study, students conducting it tried to reduce the complexity of a classifier to reduce battery drain. The changes unfortunately made the app crash and after reverting them the error persisted.

Hey Ho, Let's Go! [Conduction]

Having thoroughly planned and tested the user study, data collection can finally start. At this point, we need to instruct the participating users consistently and appropriately.

> In a user study we wanted to find out how users working with an interactive system liked its interface. However, the instruction had been interpreted differently by different people: *"While we asked participants to provide feedback concerning the interface, some commented on the task instead."* [3]

In the worst case, inaccurate wording may ruin your data.

> In a workplace study, a survey app was used to track what people did (working at computer, have a meet-

P-Hacking (Don't Do It!)

By Testing Without Hypothesis: Using the same experiment to generate and prove a hypothesis trivially yields a positive result due to circular reasoning.

Through Multiple Hypotheses: Checking large numbers of hypotheses using a single dataset greatly increases the chance of statistically significant false positives.

Through Multiple Measures: Instead of testing multiple conditions and fishing out the one that is statistically significant, we can do the same for different statistical tests.

By Iterative Testing: Adding data points one-by-one may lead to a result with statistical significance after some time. However, this may just be a local minimum.

By Removing Outliers: "Overcleaning" data until a variant of the dataset is constructed that shows the intended result.

ing, etc.). At certain times the user was asked "what are you doing next?", but what researchers actually wanted to know was "What are you doing now?"

The way how the instruction is given affects data quality. Users also need some context when participating in a study. For example, to fully answer a research question, we might collect more data than for design testing. This might affect, e.g., users' motivation of filling in questionnaires or responding to notifications in experience sampling studies.

In the participatory sensing study described above [5], some student instructors did not give the appropriate context of a research study. Thus, some participants did not like filling in questionnaires for 15 minutes after testing an app after only five minutes.

Within a user study investigating trigger types for ESM questionnaires, we were facing an unexpected low response rate of below 50% even though the questionnaire was rather short and prompted at most 14 times per day. Studies with longer questionnaires had response rates between 90 and 100%. In a final feedback round we asked for reasons. Almost all subjects reported that they did not see much sense in reacting to the prompts and answering the questionnaires. They were instructed to "react to prompts and answer them whenever they show up", but not why. Apparently, our study instructors had forgotten to explain the scenario, i.e. to point out that we were seeking insights about social activities at specific locations and location changes which require user feedback. Thus, after three weeks of user study, we ended up with little, almost unusable data.

Again, you need to consider the wants and needs of your study participants and the disruptions your study imposes.

In a field study with paragliding pilots [17], we measured in-flight using a wearable and a smartphone as data logger. Since flying conditions looked great, pilots were eager to take off. Partly because it was not communicated well by the instructor and partly out of selfishness on the pilots' side who just wanted to get into the air, some of them took off without being fully instrumented, leaving the smartphone behind. As a result, the flight data was not collected.

An important part of study conduction is also data annotation and experiment documentation. While at the time, you might think "*I can remember this...*", experience shows that you probably cannot. So write everything down, take pictures or videos (with consent!), draw sketches, etc.

Analyze This! [Analysis]

This section does not aim at describing statistical analyses in detail (for that see e.g. [14]). Instead, we will reiterate on some of the points made before. In contrast to common belief, data does not speak for itself. It is analyzed, interpreted and presented, and in each of these steps we may introduce errors, often inadvertently, As we have seen from examples above, there is a fine line between legitimate data cleaning and *p-hacking*. An important point in the analysis is to correct your p-values (respectively your significance levels) in case you are testing multiple hypotheses. At any rate, you should be clear on your methodology. When in doubt, consult a statistician or psychologist. They are really good at this. That being said, this may not prevent failure. Working with interdisciplinary, international or inter-cultural teams has its very own challenges.

Some Entertaining and Educational Resources

Spurious Correlations [20] This website (www.tylervigen.com) displays funny graphs of randomly highly correlated data and is a nice link to give to someone who cannot tell correlation from causality.

Is Most Published Research Wrong? [15] This episode (www.youtube.com/watch?v=42QuXLucH3Q) from the YouTube channel *Veritasium* very nicely covers the concepts of *p-hacking*, *publication bias*, etc.

Study Checklists Helpful resources can be found on different sites online, e.g. https://www.nngroup.com/articles/usability-test-checklist/ or https://www.cxpartners.co.uk/our-thinking/user-research-checklist/.

In a project between computer scientists and psychologists [6], the psychologists replied that they could not use our analysis when we sent it to them. After some discussion, we recognized that we had done the exact analysis they wanted. The joint project almost failed because of wording: Our *accuracy* was their *hit rate*, and our *recall* their *sensitivity*. In addition, they had another measure called *specificity* that was unknown to us and they did not know about our *precision* measure.

Discussion

When collecting the failures for this paper and writing it, two things struck us as being noteworthy. First, we noticed that for most of the described anecdotes, the failure was not reported as part of the respective paper. Even though we are convinced that the failures did not affect the validity of the eventually reported results, we still refrained from including them. The reasons for this range from adhering to page limits over not deeming it important to avoiding the perception that something is wrong with our research methodology. We should discuss our publishing culture and create more venues and incentives that encourage reporting failures.

Secondly, we felt it is important to think about how to improve the quality of research in Computer Science. In Psychology education, usually multiple semesters are devoted to studies alone, both in theory and in practice. Beyond writing this paper and teaching an introductory class about experiment design for CS students, there are many helpful external resources. For example, there are some (preliminary) guidelines about study design and analysis [11] and also recommendations about how to improve research quality [10]. We feel that CS students whose work involves user studies should be better taught to plan and conduct them.

Conclusion and Lessons

Designing and conducting user studies is a complex task. We tried in this paper to briefly introduce some of the many pitfalls that exist in designing and conducting studies and give some pointers concerning design choices and trade-offs. From conducting user studies with different methods, technologies and in different application areas, we learned:

- Be pedantic, plan well
- Always test your study small before going big
- Avoid last-minute changes

And should anything still go differently than you expected: Please report your failures. Others may learn from them.

Acknowledgements

Some examples were shamelessly taken from a talk by D. Muller in his educational YouTube channel *Veritasium* [15].

REFERENCES

1. Anja Bachmann, Christoph Klebsattel, Matthias Budde, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. How to Use Smartphones for Less Obtrusive Ambulatory Mood Assessment and Mood Recognition. In *Ubicomp'15 Adjunct*. ACM, 693–702.
2. John Bohannon. 2015. I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How. *Gizmodo*. `http://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800`. (May 2015). Accessed on July 5th, 2017.
3. Julio Borges, Matthias Budde, Oleg Peters, Till Riedel, and Michael Beigl. 2016a. Towards two-tier citizen sensing. In *Smart Cities Conference (ISC2)*. IEEE, 1–4.

4. Julio Borges, Matthias Budde, Oleg Peters, Till Riedel, Andrea Schankin, and Michael Beigl. 2016b. EstaVis: A Real-World Interactive Platform for Crowdsourced Visual Urban Analytics. In *Urb-IoT '16*. ACM, 65–70.
5. Matthias Budde, Andrea Schankin, Julien Hoffmann, Marcel Danz, Till Riedel, and Michael Beigl. 2017. Participatory Sensing or Participatory Nonsense? – Mitigating the Effect of Human Error on Data Quality in Citizen Science. *IMWUT* 1, 3 (2017).
6. Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A Wearable System for Mood Assessment Considering Smartphone Features and Data from Mobile ECGs. In *UbiComp '16*. 1153–1161.
7. F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39 (2007).
8. L. Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods* 35, 3 (2003), 379–383.
9. Niels Henze. 2011. Analysis of User Studies at MobileHCI 2011. `http://nhenze.net/?p=865`. (2011).
10. Magne Jørgensen, Tore Dybå, Knut Liestøl, and Dag IK Sjøberg. 2016. Incorrect results in software engineering experiments: How to improve research practices. *Journal of Systems and Software* 116 (2016), 133–145.
11. Barbara A Kitchenham, Shari Lawrence Pfleeger, Lesley M Pickard, Peter W Jones, David C. Hoaglin, Khaled El Emam, and Jarrett Rosenberg. 2002. Preliminary guidelines for empirical research in software engineering. *IEEE TSE* 28, 8 (2002).
12. Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *CHI '08*. ACM, 453–456.
13. Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.
14. Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.
15. Derek Muller. 2016. Is Most Published Research Wrong? *Veritasium (YouTube channel)*. `https://www.youtube.com/watch?v=42QuXLucH3Q`. (11 August 2016). Accessed on July 5th, 2017.
16. Donald A Norman and Pieter Jan Stappers. 2016. DesignX: Complex Sociotechnical Systems. *She Ji: The Journal of Design, Economics, and Innovation* 1, 2 (2016), 83–106.
17. Erik Pescara and Jonathan Gräser. 2017. Introducing A Spatiotemporal Tactile Variometer To Leverage Thermal Updrafts. In *Ubicomp'17 Adj, (Ubimount WS)*.
18. Andrea Schankin and Edmund Wascher. 2008. Unvoluntary attentional capture in change blindness. *Psychophysiology* 45, 5 (2008), 742–750.
19. Edison Thomaz, Irfan Essa, and Gregory D. Abowd. 2015. A Practical Approach for Recognizing Eating Moments with Wrist-mounted Inertial Sensing. In *UbiComp '15*. ACM, 1029–1040.
20. Tyler Vigen. 2014. Spurious Correlations. `http://tylervigen.com/spurious-correlations`. (2014). Accessed on July 5th, 2017.
21. Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.