A Prototypical Photo Sorting Study Design for Comparing Interaction Styles

Jürgen Hahn
University of Regensburg
Regensburg, Germany
juergen.hahn@ur.de

Raphael Wimmer
University of Regensburg
Regensburg, Germany
raphael.wimmer@ur.de

ABSTRACT

We discuss a first version of a extensible study design for a generic image-sorting task. It allows for comparing qualitative and quantitative properties of different interaction styles (e.g., direct manipulation, CLI, tangible interaction), input modalities (e.g., mouse vs. touch screen), output modalities, and UI implementations. Therefore, the study design may be of use for designing reproducible and replicable studies. Study participants are asked to sort a set of 27 photos into five categories where each of the photos depict one distinct topic belonging to only one of the categories. We conducted a first pilot study using this design, comparing a desktop GUI, an interactive tabletop, and physical photos. Task completion time was significantly lower when sorting physical photos than in the other two conditions. The study results may serve as a baseline and show limitations of the preliminary design.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Usability testing**; **HCI theory, concepts and models**.

KEYWORDS

sorting task, user study, user interfaces, digital vs. physical

ACM Reference Format:
Jürgen Hahn and Raphael Wimmer. 2019. A Prototypical Photo Sorting Study Design for Comparing Interaction Styles. In *Mensch und Computer 2019 (MuC '19), September 8–11, 2019, Hamburg, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340764.3344892

MuC '19, September 8–11, 2019, Hamburg, Germany

ACM ISBN 978-1-4503-7198-8/19/09…$15.00
https://doi.org/10.1145/3340764.3344892

Figure 1: In the generic study design that is presented in this paper, users have to sort 27 images into five categories: city, vacation, food, pet, screenshot (from left to right).

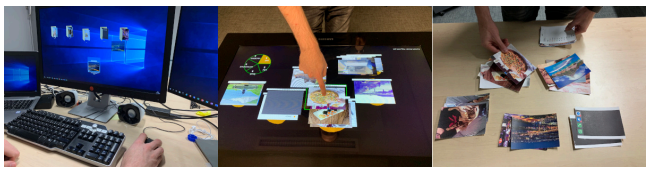

Figure 2: The study setups for image sorting on a desktop computer, on an interactive table or physical from left to right.

1 MOTIVATION

Choosing the right study design and tasks for a user study is non-trivial but essential. When evaluating interaction styles or techniques, user interfaces, or devices, lab studies are a standard approach. They achieve high internal validity as tasks and environment are strictly controlled. Due to these intentional limitations, the results of such studies do not necessarily directly transfer to real world usage, however.
For an ongoing research project, we are developing study designs and sets of standardized tasks that allow for qualitative and quantitative evaluation of different interaction styles, interaction technique, graphical representations, and

input/output devices. Such a study design's goal is to allow, e.g., comparing the performance of a novel input method to that of mouse input for non-trivial tasks that are representative of real-world use cases. The study design should facilitate reproducibility and allow for replicable results. Also, the design should reduce the probability of errors, while also mitigating effects of confounding variables. The results of different instances of the study should be comparable to each other, e.g. allowing to track the progress of a novel UI under construction.

In this paper we propose an initial prototypical study design for an image sorting task ẃith the goal to compare different interaction styles and input modalities. We conducted an initial test of the study design wherein 22 participants sorted 27 images into five pre-defined categories (Figure 1) using three different interaction styles and input modalities: direct manipulation using mouse or touch input as well as manual interaction with physical photos (Figure 2).

2 RELATED WORK

Only few standardized study designs exist in HCI research. Most commonly, standardized tests are used for measuring text entry performance or pointing performance. For measuring text entry performance, multiple researchers have used the *TEMA* system [3] in the past [4]. Pointing performance is usually evaluated using standardized tasks based on ISO/TC 9241-411 [6] or inspired by task designs used in research on Fitts' Law [5] and the Steering Law [1]. However, these methods evaluate very simple, formalized tasks which work very well for their respective contexts but are much simpler than real-world interactions.

Sorting tasks have been used in the past to compare novel interaction styles such as tangible interaction and input modalities (such as touch interaction) to traditional ones. However, those tasks are rarely embedded into reproducible study designs which make it hard to replicate results or compare the results to those from later studies. For example, Terrenghi et al. [7] investigate qualitative and quantitative differences between multitouch interaction and tangible interaction. The study participants completed two tasks (completing a puzzle and sorting photos into three categories) - once on an interactive tabletop and once using physical tiles/photos. Participants brought their own, most recent and unsorted images for the sorting task, and sorted these photos into three categories of their own choosing. Therefore, the categories themselves and the topics of the images are highly subjective and prone to deviate heavily from participant to participant. This means that the results of their study are not comparable across participants and to other potential studies of the same design. Using a study design with a set of standardized tasks and also utilising a defined and valid set of auxiliary artifacts, such as photos in this case, would have allowed results to be compared within the study and across reproductions and replications.

Carvalho et al. [2] employ a selection task with four input modalities in order to explore the relationship between different interaction paradigms or styles and user groups. For selection via mouse or touch, participants had to select one of eight buttons in a graphical user interface. In order to perform selection with tangible objects whose sides were covered with different symbols, participants had to hold them in front of a camera. Afterwards, participants had to activate a button using gestures and had to move a crosshair on top of the button by moving their arm. However, as the *tangible interaction* condition required more steps than the mouse or touch counterparts, the reported performance measures actually do not reflect genuine differences between these interaction styles. Using a study design with standardized tasks, where all tasks use the same artifacts and no task requires additional interactions, would have allowed for actually comparing interaction styles, input modalities and, user interfaces.

3 STUDY DESIGN

Overall, the reproducibility and ability to replicate studies in the field of HCI is commonly under-developed or even possibly under-appreciated. Therefore, we propose a study design using an image-sorting task that can be used for characterizing and comparing the performance of many different interaction styles, interaction technique, graphical representations, and input/output devices. Our goal is to maximize reproducibility of the study design and replicability of results.

3.1 General Overview

In the proposed study design, participants need to repeatedly sort 27 images into five categories. In each condition, they use a different combination of interaction style, interaction technique, graphical representation, and input/output device. We chose an image-sorting task, because the interactions it requires are representative of many typical tasks in everyday life and all sorting task variations share the same abstract characteristics:

Users view a property of an object, select it, decide on a location, and move it to that location. They release the object and repeat the process until all objects are sorted. Occasionally, users may change their mind and move an object from one sorting location to another one, e.g. because of an error they made or change of sorting context.

We let participants sort images because these are fast-to-process, cross-cultural, language-independent and common objects. Because a sorting task has a common set of actions (e.g. view, create, select, move), it is suitable for evaluating

many interaction styles and devices, and allows for recording quantitative measures (e.g., task completion time, error rate, mental workload) and qualitative observations (strategies, usability problems, etc.). Furthermore, with regard to interaction styles, the sorting task is suitable for e.g. CLI, direct manipulation, dialogue, physical interaction and with regard to devices it is suitable for e.g., mouse, touchscreen, keyboard, small and large displays, or VR/AR headsets.

3.2 Study Setup, Task and Metrics

For the general study setup, we suggest a within-subjects design in which every participant performs the same sorting tasks using multiple interaction styles, modalities, user interfaces, or devices (or a combination thereof). In order to cancel out learning effects, conditions are counter-balanced using the balanced latin square method. Every participant is shown an exemplary correct sorting of the images before the test starts. This avoids misunderstandings and allows participants to familiarize themselves with the images and their correct classification. In case of evaluating particular unfamiliar interaction styles, modalities or devices, participants are given time for familiarization before the actual experiment begins. For each interaction style or modality the same set of images is used. These must be displayed in an equivalent way to the participant (Figure 2). Participants select the topmost photo from a single pile of unsorted images and assign the photo to a category using the modality required by the currently tested condition. They repeat the process until all images are sorted. Task completion times and error rates are logged. The time required for setting up the categories (e.g. creating, renaming and re-positioning of target folders on a desktop GUI) needs to be measured and reported separately from the time required for sorting. Participants are observed and recorded during the test, in order to document their sorting strategies and their behaviour when making an error and fixing it. Thinking-aloud should be reserved for separate test runs where no quantitative data is collected.

3.3 Choosing an Image Set

We procured 31 images available under a CC-0 license or from a personal set of *pet* images, belonging to exactly one of the following categories: *vacation*, *city*, *food*, *pet* and *screenshot*. These categories share little to no characteristics and minimize the required mental workload for users to classify the images' categories. In a preliminary study, we condensed the image set to only contain images that could be unambiguously assigned to exactly one category. Four participants (all male; mean age: 28.75 years old; st.dev.: 2.48) assigned the images to one of the five categories or an *unsure* category. We only kept images that have no ambiguity in regard of their interpretation (100% inter-rater reliability). The final image set consists of 27 images: seven pets, seven food items, seven vacation locations, three aerial photographs of cities and three smartphone screenshots (Figure 1). We discarded images that either showed ambiguity in terms of whether they are most likely a city or a vacation location or were classified as *unsure*.

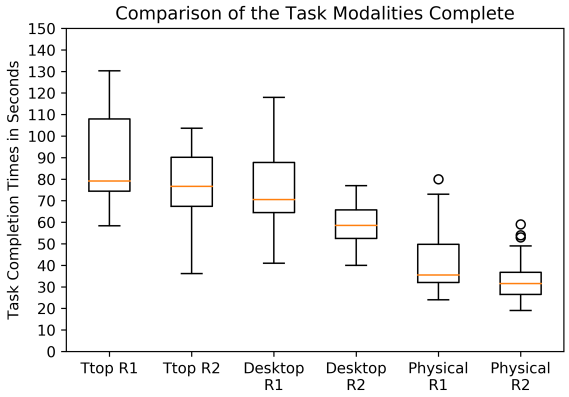

Figure 3: Comparison of the task modalities for both repetitions.

4 EVALUATION OF THE STUDY DESIGN

In order to refine the study design, we conducted an initial study. We recruited 22 participants (16 male and 6 female) whose age ranged from 19 to 56 years old (mean: 25.59 y/o, st.dev.: 9.59). 20 participants were right-handed. The participants of the preliminary study did not take part in this study. The experiment was conducted in a standard lab environment. At the beginning, each participant received a demonstration of the correct sorting of the images in physical form and was given up to 10 minutes for familiarizing themselves with an interactive tabletop (Samsung SUR40). The participants performed the sorting task with the set of photos using three different interaction styles and modalities: direct manipulation via mouse (desktop computer) and touch (interactive tabletop), and manual interaction with physical photos. The sequence was repeated once (**R1** and **R2**). Due to counter-balancing, every 6th participant performed the same sequence. Participants were seated in front of the screen of a laptop workstation with access to keyboard and mouse (Figure 2). For physical sorting, the participants were seated at a table . The participants stood in front of the interactive tabletop (Figure 2). The images were stacked as a pile, whereby the topmost image was big enough to be fully visible and at the same location in front of the participant within each condition. Figures 3, 4 and 5 show the comparison of the task completion times per interaction style and modality per task repetition (R1, R2).

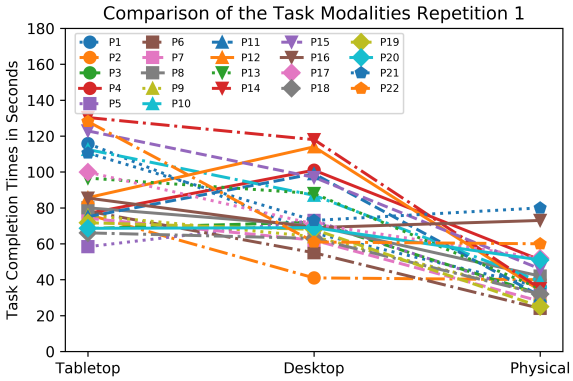

Figure 4: Comparison of the task modalities for Repetition 1.

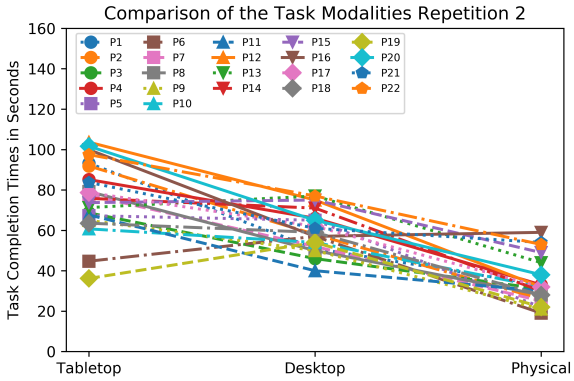

Figure 5: Comparison of the task modalities for Repetition 2.

We calculated two-sided t-tests with Bonferroni correction for related or repeated samples for our task completion time distributions per style and modality and repetition. The physical sorting is always performed significantly faster than the desktop equivalent (R1: $p < 0.0001$; R2: $p < 0.0001$). However, the desktop sorting task is not significantly faster compared to the interactive tabletop, and only faster in R2 (R1: $p < 0.4825$; R2: $p < 0.0037$). Sorting physical photos is significantly faster than sorting on the interactive tabletop (R1: $p < 0.0001$; R2: $p < 0.0001$). Learning effects were observed for all three conditions (Figures 3 and 5). As expected the participants made few errors. The total number of errors made decreased from the first iteration to the second one. Multiple participants mentioned that the interactive tabletop's latency was annoying. When using the desktop GUI, participants who made errors sometimes did intentionally not correct them and argued that this would be too much of a hassle. Because of this, the only occurrences of incorrect sorting after task completion were on the desktop UI. The physical sorting task was preferred by most as it felt the most natural, fastest and smoothest and most forgiving in terms of error correction.

5 DISCUSSION AND LIMITATIONS

This papers goal was to investigate whether the proposed study designs works at all and can be used for different interaction styles and input modalities. We did not observe any inherent problems with the study design itself, the chosen image-sorting task or during the conduction of the study which suggests that we achieved this goal However, there are also limitations to our overall approach and proposed general study design for photo sorting. Foremost, we did not replicate the study. Further, the photo sorting task is a purely visual task. Therefore, a comparison to non-visual tasks, such as audio-only, tactual or haptic, is limited. However, arguably most of the common interaction styles in HCI are visual. With regard to our study design, we observed that two repetitions appear not to be enough for learning effects to converge in terms of task completion times. Therefore, with more repetitions, participants would reach their personal optimum and so more realistic measurements of task completion times occur, further improving the validity of the results.

6 CONCLUSION AND FUTURE WORK

The results from the current study can serve as a baseline for evaluating other interaction styles, techniques, user interfaces and devices. The next step is to replicate the study, preferably by an independent party and investigate whether the results are the same. Further, we intend to conduct focus groups across multiple user groups, in order to establish an intersecting set of tasks across users' respective profession domains, such as administration, arts, medicine, etc. We intend to explore the set of task in these domains and then attempt to compare them across all domains. The goal is to reach a generic or standardized set of tasks and study designs for evaluating interaction styles and techniques, input modalities and user interfaces in general. There, we plan to investigate whether visual tasks can be compared to non-visual tasks.
A detailed description of the study setup and all materials required for reproducing the study can be found at:
https://hci.ur.de/projects/standardized_studies

ACKNOWLEDGMENTS

This project is funded by the Bavarian State Ministry of Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

REFERENCES

[1] Johnny Accot and Shumin Zhai. 1997. Beyond Fitts' Law: Models for Trajectory-based HCI Tasks.In *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '97), 250–250. https://doi.org/10.1145/1120212.1120376

[2] Diana Carvalho, Maximino Bessa, and Luís Magalhães. 2014. Different Interaction Paradigms for Different User Groups: An Evaluation Regarding Content Selection.In *Proceedings of the XV International Conference on Human Computer Interaction* (Interacción '14), 40:1–40:6. https://doi.org/10.1145/2662253.2662293

[3] Steven J. Castellucci and I. Scott MacKenzie. 2011. Text Entry Metrics for Android.*Text Entry Metrics for Android*. Retrieved from http://www.cse.yorku.ca/~stevenc/tema/

[4] Steven J. Castellucci and I. Scott MacKenzie. 2011. Gathering Text Entry Metrics on Android Devices.In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '11), 1507–1512. https://doi.org/10.1145/1979742.1979799

[5] Paul M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement.*Journal of Experimental Psychology* 47, 6: 381–391. https://doi.org/10.1037/h0055392

[6] International Organization for Standardization. 2012. Ergonomics of human-system interaction - Part 411: Evaluation methods for the design of physical input devices (ISO/TS 9241-411:2012). Retrieved July 19, 2019 from https://www.iso.org/standard/54106.html

[7] Lucia Terrenghi, David Kirk, Hendrik Richter, Sebastian Krämer, Otmar Hilliges, and Andreas Butz. 2008. Physical Handles at the Interactive Surface: Exploring Tangibility and Its Benefits.In *Proceedings of the Working Conference on Advanced Visual Interfaces* (AVI '08), 138–145. https://doi.org/10.1145/1385569.1385593