# Evaluating Peripheral Interaction

**Doris Hausen**[1], **Aurélien Tabard**[1,2], **Attila von Thermann**[1], **Kerstin Holzner**[1], **Andreas Butz**[1]
[1]University of Munich (LMU), HCI Group, Amalienstr. 17, 80333 Munich, Germany
[2]LIRIS - Université Lyon 1 & CNRS UMR5205, 69622 Villeurbanne, France
doris.hausen@ifi.lmu.de, aurelien.tabard@univ-lyon1.fr, thermann@cip.ifi.lmu.de,
holznerk@cip.ifi.lmu.de, andreas.butz@ifi.lmu.de

**ABSTRACT**
Peripheral interaction, like ambient information systems (AIS), aims at leveraging the periphery of our attention. While ambient information systems address the perception of information, peripheral interaction targets lightweight interaction outside of the current focus of attention. A number of prototypes have demonstrated the value of peripheral interaction through long-term in-situ deployments. Such studies are particularly suited to evaluate peripheral interaction since they enable the integration of devices into daily routines and thereby move interaction to the periphery of attention. However, they do not lend themselves well to early design phases. In fact, the design process completely lacks early evaluation tools to assess design choices.

We propose an experimental method for the evaluation of peripheral interaction in early design phases. In a case study, we compared the results of an eight-week in-situ deployment with the results of this laboratory experiment. We carried out the study with both, novice and experienced users (who had participated in the in-situ), and found comparable results across all three situations (in-situ and lab with novice and experienced users).

**Author Keywords**
Peripheral Interaction; Evaluation

**ACM Classification Keywords**
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

**General Terms**
Human Factors; Design; Measurement.

**INTRODUCTION**
When interacting with the physical world, we carry out activities in parallel with no or only minimal attention. We drink while reading, we walk while talking and we sing along a song while preparing dinner. In contrast to these everyday activities, digital devices are all-too-often requiring undivided attention. This all-or-nothing approach leads to frequent context switches (e.g., switching between applications), which disrupts users from their primary task.


*TEI'14,* February 16 - 19 2014, Munich, Germany

http://dx.doi.org/10.1145/2540930.2540941

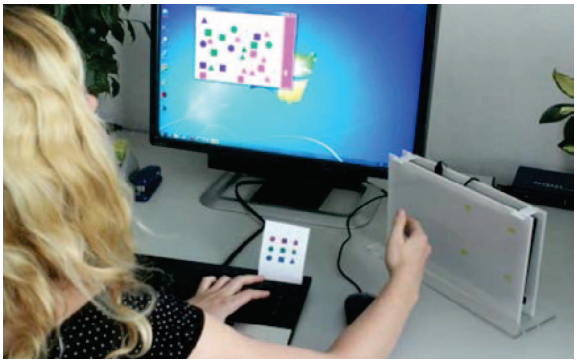

**Figure 1. User interacting with the music controller (here freehand interaction) while carrying out the primary task.**

Peripheral interaction tries to overcome this by moving (especially small) tasks to the periphery of attention. In the tradition of calm technology [26], we aim at simple and casual interactions in the periphery of attention, relying on human capabilities such as proprioception [6], divided attention [14,27], and habitual processes [2] that can be carried out with minimal conscious control. Similar to ambient information [21], our goal is to only cause minimal distraction, but in contrast, peripheral interaction does not only aim at displaying information but also at acting on information in the periphery. Peripheral interaction is normally used for small side tasks (e.g., changing the instant messaging status) or supportive tasks (e.g., changing the size of the brush while drawing in a graphics editing program) in parallel to a larger main task (e.g., reading, writing, drawing).

This paper discusses evaluation methods for peripheral interaction systems. Most systems in the literature have been evaluated through field studies [3,4,9,10,11]. This is motivated by the learning time needed to push an interaction from focus to periphery. However, this also leads to an evaluation gap in the early research phases, since field studies require a fully functional (i.e., late) prototype. Usability problems are only discovered when the field study is already running and affect the results. Hence we analyzed the requirements for successfully evaluating peripheral interaction in a lab setting and tested our methodology by comparing the results from a field evaluation to the results from the lab. The lab study itself (see Figure 1) was run twice, once with participants who had already taken part in the in-situ deployment and were familiar with the system, and once with participants, who did not know the system beforehand.

## BACKGROUND ON EVALUATION

Peripheral interaction is a new interaction paradigm, which emerged in the last few years. With every new type of interaction the question of a suitable design process and evaluation method arises. The iterative design process of regular (non-peripheral) systems ideally includes prototypes of various degrees of fidelity and the corresponding evaluation methods (see Figure 2). Low fidelity prototypes, such as paper prototypes and sketches are used at early stages [5] with methods such as cognitive walk-throughs or as inspirations for focus groups. These early evaluations help to decide between different designs and to discover conceptual usability problems. However, because of their low fidelity, early prototypes are often more attention demanding than a polished system. This makes it difficult to gain insights on the intrinsic quality of a peripheral device where interaction must fall into the periphery of attention, which usually can only be expected at the last stages of prototyping [18]. Later development stages use working prototypes at higher fidelity and move to more empirically solid evaluation methods, such as lab studies or field deployments [8]. Currently, in-situ deployments are the usual choice for evaluation of peripheral interaction lacking feedback in early design stages.

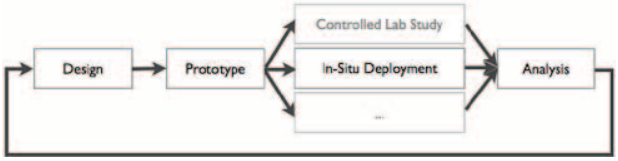


**Figure 2 – The iterative design process is run through several times for different development stages. For peripheral interaction usually only in-situ deployments are carried out (black boxes).**

### Evaluation of Peripheral Interaction

Up to now, most peripheral interaction systems have been tested with in-situ deployment.

#### *In-Situ Deployments*

Researchers motivate their decision to carry out in-situ studies by the need to evaluate their peripheral systems in the actual context of use [3], which is required for the interaction to become a routine and shift to the periphery [4]. They refer to findings from AIS [3,10], which also rely on the real world contexts for evaluations [12]. The durations of these in-situ deployments differ between two and eight weeks per participant. The number of participants also ranges between two and eight. Most of these studies rely on observations [3,4], questionnaires [3,10,11] and (semi-structured) interviews [3,4,9,10,11] . Furthermore the usage of the device is usually logged [3,9,10,11], when possible this usage is compared with log data without the peripheral device [10].

#### *Controlled Lab Experiments*

The only work in the context of peripheral interaction that has been evaluated with a lab study is PolyTags by Olivera et al. [20]. To mimic the dual-task situation participants had to count the occurrences of a specific vowel in a text. While counting, participants were interrupted to carry out interactions with the PolyTag prototype. Olivera et al. measured the error rate (in relation to the number of words) and the performance (words per second). They found that participants made significantly fewer errors and were significantly more efficient with PolyTags than with a traditional GUI.

### Evaluation in Related Fields

Similar to peripheral interaction, AIS[1] targets the periphery of attention [21]. Most researchers designing peripheral interaction thus base their evaluation on findings from ambient information. In both fields researchers deal with the fact that the task presented as less important is most important for the evaluation [15]. Additionally peripheral interaction deals with dual-task situations, thus multitasking research is also a source of inspiration as it aims to uncover supporting strategies for interruption management [1,16].

#### *In-Situ Deployments*

Hazlewood et al. [12] argue that ambient information needs to be integrated into the everyday life to work properly, which is hard to simulate in the lab. Indeed the combination of "out-of-the-ordinary" systems and researchers observing the user steer the users' attention in an unnatural way [12]. Hazlewood et al. thus tried to eliminate direct observation in two case studies. Experience sampling [13] or scheduled interviews [23] are alternatives to eliminate direct observations. Generally field studies pose problems such as uncontrolled variables and events [12] and finding users that fit the context of the prototype [24]. Furthermore, as many ambient systems do not aim at a clearly defined task, usage is often ambiguous. For gathering data, sometimes artificial events [12] are used. In addition, privacy needs have to be addressed, especially when testing a prototype in participants' homes [24].

#### *Controlled Lab Experiments*

Lab studies in the context of ambient information address awareness, distraction, learnability, comprehension, aesthetics, suitability and flexibility [17,22]. Usually dual-task studies are used to distract users and move their attention away from the ambient display and the secondary task. Distraction tasks include mathematical tasks (calculating, counting) [1,13], comprehension tasks (reading a text or news, analyzing a graph, quizzes) [1,17], interacting with typical interfaces (registration forms, email sorting) [1,16,17] and click tasks [22]. Also very complex tasks are used where participants are asked to read emails, acquire information and reply to the email [7]. Task completion time and error rate are logged for the primary task [1] but emotional distress is also of interest, as interruptions are known to cause annoyance and anxiety [1,16]. Performance in the secondary task, especially for ambient information systems, is often measured through questions about the content, asked at the end of the study [13,22].

#### *Analytical Methods*

An additional evaluation method is Mankoff et al.'s collection of Heuristics for Ambient Displays [15]. Further at-

[1] We use the term "Ambient Information System" as proposed by Pousman et al. [21] including peripheral displays and notification systems etc.

tempts to evaluate peripheral or notification displays include the IRC model proposed by McCrickard et al. [20]. IRC stands for interruption, reaction and comprehension and offers a way to classify them.

## DESIGNING A CONTROLLED LAB EVALUATION

The evaluation of peripheral interaction requires at least two tasks: A primary task, which should be the focus of the participant's attention, and a secondary task, which should be carried out in the periphery. This secondary task is usually a given: the task supported by the peripheral system being evaluated.

### Designing the Primary Task

In a dual-task study the conditions of the primary task will have an impact on the measures of the secondary task. While the primary task should not only be related to a real life situation, it should also be abstracted in order to control users' attention and steer it away from the secondary task (i.e., the peripheral system). We analyzed the properties of primary tasks and propose here a list of parameters, which should be defined for any primary tasks. These parameters can be fixed throughout the experiment or vary according to the elements being tested in the secondary task:

*Input Channel:* The input channel(s) used to carry out the primary task (e.g., mouse, keyboard, body movement, combinations).

*Output Channel:* The output channels are used or "blocked" by the primary task (e.g., auditory, visual and/or haptic channels are typical for desktop PC based tasks).

*Input Interruptibility:* The degree of continuous input in the primary task. Low input interruptibility: Participants must constantly attend to the primary task. High interruptibility: Participants can take breaks from the primary task at any time to carry out other actions.

*Attentional Interruptibility:* The degree of continuous attention required to execute the primary task. Low attentional interruptibility: A shift of attention has large detrimental effect on primary task performance. High attentional interruptibility: A shift of attention has no or marginal negative effect on primary task performance.

We abstracted the dimensions of the primary tasks so that they could be reused. Input and attentional interruptibility together define the difficulty of a task, thus the number of mental resources required to carry out the primary task (cf., Kahnemann [14]) can differ for different manifestations of interruptibility. In practice, like with most controlled studies, evaluators should take extra elements into account such as experience or motivational factors. The parameters should be adjusted to the secondary task and consider elements from the field to improve external validity. On a practical level, a large variety of primary tasks can be considered for testing peripheral devices. We here focus on a desktop computer scenario (thus not using Olivera et al.'s [20] vowel counting task, which did not use mouse and keyboard as input channel) as this is still very common to interact with digital data and many side tasks can be imagined (e.g., controlling music or monitoring the instant messenger status).

### Comparing Event-Based and Continuous Primary Tasks

The *input* and *output channel* parameters are mostly defined by the use case of the secondary task and usually constant throughout the experiment. The *input* and *attentional interruptibility* parameters are harder to define and control. To compare alternatives in terms of interruptibility, we conducted a preliminary experiment comparing *event-based* and *continuous tasks*. We based our experiment on SquareClick [22], which displays a black square that changes its location. Users need to click the square after a location change within one second.

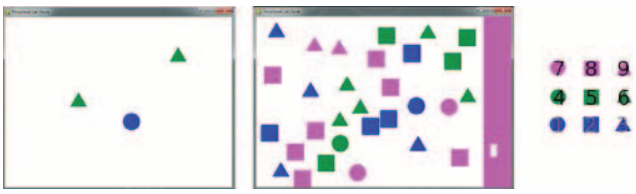


**Figure 3. Left: Event-Based Task: Shapes appear randomly and should be removed; Center: Continuous Task: Shapes of a given color should be removed (here pink); Right: Shape/color combinations mapped to the number-pad. The corresponding number should be pressed while clicking on the shape.**

#### *Event-Based Primary Task*

The first task is event-based, meaning *low attentional interruptibility* and *low input interruptibility*. Similar to SquareClick, users have to react to events, but we modified the task to have several items of different shape (square, triangle, circle) and color (green, blue, pink) appearing in a randomized order at randomized locations (see Figure 3 left). Instead of clicking on an item, participants have to input a number on the keypad corresponding to the specific combination of shape and color (see Figure 3 right).

**Preliminary Study:** We presented the event-based task in parallel to an artificial peripheral task to four members of our lab. After carrying out the dual-task study (they did not know that our actual focus was the primary task) we invited them to a group discussion to get feedback on the primary task. We identified two problems: (1) The task was not engaging but perceived as tiring and monotonous. Gaps without anything to do appeared between different events and no motivational feedback was offered (e.g., a counter of successfully clicked items). (2) Participants could not develop strategies and were either especially stressed or bored because of the event-based nature of the task. We therefore opted for a second, continuous primary task, which is also more in line with general tasks performed at a desktop PC, where interactions usually are not that time critical.

#### *Continuous Primary Task*

The second task is based on continuous input organized into rounds, thus meaning *high input* and *high attentional interruptibility*. Each round consists of 30 items of different shapes and colors (cf., event-based task) appearing at random locations on the screen. Participants must now click

and remove all items of the color displayed in the sidebar of the window (see Figure 3 middle) while inputting the correct number for the color/shape combination on the number pad, thus controlling the input channel and enforcing bimanual interaction. When all items of one color are deleted, a new round starts immediately, new items and a new color in the sidebar appear. Additionally we included a counter showing the number of removed items. This task makes room for different strategies for the primary task itself, for example removing all items with the same shape (e.g., all squares) and therefore moving the mouse over greater distances but not changing the selected number or clicking close-by items but therefore switching the key regularly to match the shape. To interact with the peripheral task, immediate reaction is still possible, but participants can also decide to finish all items with the same shape (i.e., one number) or even one round (which lasts 10 seconds on average).

**Preliminary Study:** To get a first understanding of the evaluation method and particularly the primary task, we carried out a study with eight participants, which were 22 years old on average. We told them that we were interested in a new peripheral device (to hide the fact that we were testing the evaluation method itself). As a peripheral device we used a previous prototype [10]. Participants were located at a table equipped with a display, keyboard and mouse as well as the peripheral device. We provided a cheatsheet (Figure 3 right) with the numbers corresponding to each item (shape and color) attached to the number-pad. Before carrying out the task, participants had a training to familiarize themselves with the primary and the peripheral task. We then asked the participants to carry out just the primary task for five minutes as baseline measurement (number of removed items, number of completed rounds, errors (wrong color/key pressed)). Afterwards a five-minute trial with both, primary as well as peripheral task was carried out. Besides logging the interaction the experimenter took notes on focus shifts, hand movements and other observations.

During our evaluation we could not shift the peripheral device to the periphery, but we found usability problems for the device. This achieved our goal to support the design process in an early stage. We further found a degradation of the performance in the primary task, when using the peripheral device. Testing different early design concepts with this method can therefore help to make informed decisions on which concept to pursue further, assuming the design with the least degradation works best.

## CASE STUDY: COMPARING THE FIELD AND LAB

To validate the initial results we carried out a controlled lab study using our method for one of our prototypes – the peripheral audio controller – which we had already studied in an eight-week in-situ deployment [11]. We recruited participants, who did not know our prototype but also invited participants from the in-situ deployment back to the lab. By comparing these two groups we expected to learn whether familiarization with the devices had a significant effect on the results and would therefore render our controlled lab study method invalid. The lab evaluation methodology was developed in parallel to building the peripheral audio controller. Findings from the in-situ deployment therefore did not influence the design of the controlled study methodology but when carrying out the lab study, we knew the results of the in-situ deployment. This offers the possibility to compare results between both studies and especially enables us to ask experienced users back to the lab.

### The Peripheral Audio Controller

The prototype was built to compare different interaction styles for peripheral control (graspable, touch and freehand interaction). As use case we chose controlling an audio player and participants could carry out simple commands (next/previous song, pause/play, volume control) while keeping their primary focus on their current task (see Figure 4). Some keyboards also provide similar functionality with media keys. Therefore we included them as a fourth condition alongside the three peripheral interaction styles. The audio controller was studied in an eight-week in-situ deployment with eight participants. We logged all interactions with the devices, and generally all interactions with the audio player. Additionally we carried out five semi-structured interviews with each participant to get exhaustive insights on usage and personal preferences. Details on the development of the prototype as well as the field study are extensively discussed in [11]. Results of the field study relevant for the comparison of results are presented alongside the results of the lab study in the following sections.

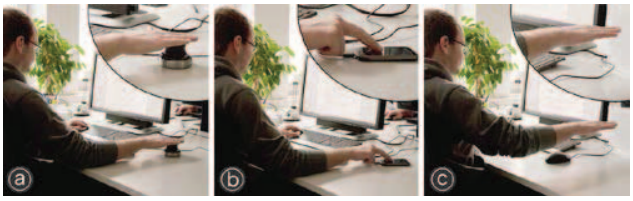


**Figure 4. Participant controlling his audio player with the peripheral devices: a) graspable, b) touch and c) freehand.**

### Lab Study of the Peripheral Music Controller

We designed the peripheral music controller to control music while carrying on another task. Listening to music is an activity involving little interaction and intrinsically motivated, consequently we used triggers, which appeared randomly during the study to push participants to interact with the peripheral music controller. To trigger the pause/play gesture we stopped the music and asked participants to start the music again. To trigger next/previous we added noise to some songs, which was unpleasant to listen to. To trigger the volume gesture we changed the volume noticeably and asked participants to keep the volume at a medium level.

#### *Procedure*

We had a mixed-model design with two independent variables (interaction style and user group). Everybody tested all interaction styles (graspable, touch, freehand, media keys) but we had two distinct user groups (experienced from the

in-situ deployment and novice users unfamiliar to the devices). We counterbalanced the order of interaction styles.

Every participant was seated at a standard desktop computer with mouse and keyboard. The peripheral device was situated on the right (which was the preferred location during the in-situ deployment, see Figure 1). After a short introduction and a questionnaire asking for demographic data and their usage of music players in parallel to other tasks we introduced the (continuous) primary task and after a training phase carried out a baseline measurement for the performance in the primary task (i.e., interacting with the primary task without interruptions by the peripheral task) for two minutes. Afterwards we introduced the first interaction style (graspable, touch, freehand or media keys), explained the respective gestures, told participants that they should reduce gazing at the device as much as possible and carried out a two minute training including the triggers. Subsequently we started the first trial with the primary and peripheral task in parallel. One round lasted five minutes and participants were instructed to react to each trigger in a reasonable time frame, i.e., participants could adapt different strategies in the primary task. Each round, including 16 triggers, was carried out for all interaction styles and followed by a questionnaire. After all four interaction styles were tested we ended the study with a questionnaire comparing the interaction styles. The last questionnaire slightly differed for experienced and novice users, asking experienced users whether it had helped them during the study that they were already familiarized with the devices beforehand. During the whole study the music player (we used iTunes) was minimized, hence participants only had auditive, i.e., functional feedback [25] about their interactions.

As dependent variable we measured the number of successfully removed shapes in the primary task and calculated the error rate (errors – wrong shape/color – in relation to the overall removed items). In the peripheral task we measured the reaction time (time between the trigger and the start of the execution of the peripheral task) for each command and all errors (wrong gestures, gestures without trigger, no reaction to trigger, tracking error of the device). Furthermore with the help of video analysis we analyzed the gazes to the device. All Likert scales used in the questionnaires ranged from 1 = "I totally disagree" to 5 = "I totally agree".

*Participants*

For our controlled experiment we invited two distinct user groups to our lab: twelve new participants who did not now the peripheral music controller and six out of the eight participants from our in-situ deployment. Our main reason for doing this was to compare whether there was a difference in results between users who were already familiar with the devices – experienced users (Exp) – and could carry out the interaction in the periphery and the typical lab study participants who were unfamiliar with the system they are testing – novice users (Nov). The average age of the experienced users was 25 years while the novice users were 22 years old on average. No participant reported any hearing impairment. Furthermore all participants stated to multitask while working on the computer, and to listen regularly to music and interact with their player in parallel to other tasks.

**Results for the Controlled Lab Experiment**

We analyzed and compared the data for both user groups and if possible also the results from the in-situ deployment

*Quantitative Data for the Primary Task*

We carried out a Two-Way Mixed ANOVA. For pair-wise post hoc tests, we used Bonferroni-corrected confidence intervals to retain comparisons against $\alpha = 0.05$. When the assumption of sphericity was violated, we used Greenhouse-Geisser to correct the degrees of freedom. All unstated $p$-values are $p > 0.05$.

**Performance:** We logged the number of correctly removed shapes in the primary task. We did not find any significant effect for *Interaction Style* × *User Group* but we found a significant effect for *Interaction Style* ($F_{3,48} = 8.453$ $p < 0.001$). Post hoc tests showed that performance in the primary task was significantly better while interacting with the *Graspable* in the periphery compared to *Freehand* and *Media Keys* ($p = 0.004$). As Figure 5a shows, the most correct shapes have been removed while using the *Graspable* followed by *Touch*, *Media Keys* and *Freehand*.

**Error Rate:** To assess the success in the primary task we also calculated the error rate (errors in relation to all clicked shapes). Statistical analysis did not show any significant effect for *Interaction Style* and *User Group*. Figure 5b gives an overview of the error rate with *Touch* provoking more errors than *Media Keys*, *Freehand* and *Graspable*.

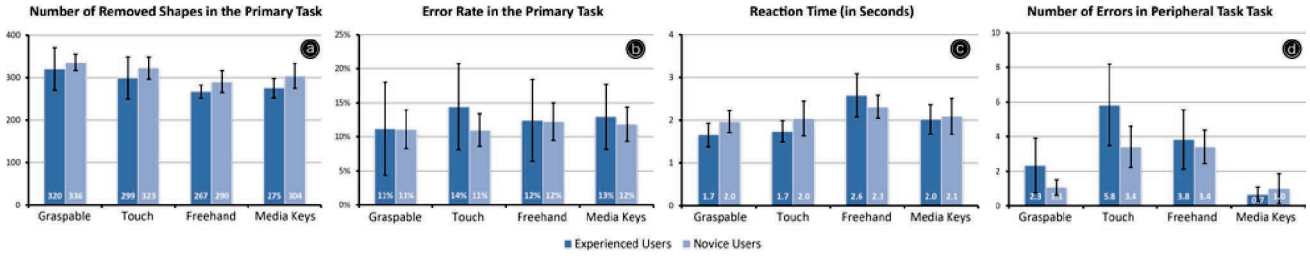


**Figure 5. a) Performance in the primary task; b) Error rate in the primary task: c) Reaction time (between trigger and interaction in the peripheral task); and d) Number of errors in the peripheral task. Numbers in bars indicate the mean value; error bars indicate 95% confidence intervals.**

*Quantitative Data for the Peripheral Task*

Again we carried out a Two-Way Mixed ANOVA.

**Reaction Time:** Reaction time is the time between the trigger and the user starting the interaction with the peripheral device. We did not find a significant effect for *Interaction Style* × *User Group* but for *Interaction Style* ($F_{3,48}$ = 8.243 $p < 0.001$). Pairwise comparison shows a significant difference for *Freehand* compared to *Touch* and *Graspable* ($p = 0.004$). As Figure 5c shows, the reaction time was shortest for *Graspable* followed by *Touch*, *Media Keys* and *Freehand*.

**Errors:** Errors include any wrong input (wrong gesture, gesture without trigger, no reaction to trigger, tracking errors). Statistical analysis did not show a significant effect for *Interaction Styles* × *User Group* but for *Interaction Style* ($F_{3,48}$ = 14.863 $p < 0.001$). Post hoc tests showed a significant difference for *Graspable* compared to *Touch* and *Freehand* ($p = 0.03$) and *Media Keys* compared to *Touch* and *Freehand* ($p = 0.003$). As depicted in Figure 5d, the least errors were carried out with *Media Keys* followed by *Graspable*, *Freehand* and *Touch*.

**Gazes:** Aiming at minimal visual attention we analyzed the number of interactions without looking at the peripheral device. We did not find a significance for *Interaction Style* × *User Group* but one for *Interaction Style* ($F_{1.821,29.133}$ = 28.540 $p < 0.001$). Pairwise comparison revealed significant effects for *Graspable* compared to *Freehand* and *Media Keys* ($p < 0.001$) and *Touch* compared to *Freehand* and *Media Keys* ($p = 0.047$). The least gazes to the peripheral device have been carried out with the *Graspable* (Exp: *m*=68,0%; *sd*=29,1%; Nov: *m*=70.5%; *sd*=22.9%) followed by *Touch* (Exp: *m*=50.6%; *sd*=39.7%; Nov: *m*=20.8%; *sd*=32.2%) and *Freehand* (Exp: *m*=10.4%; *sd*=14.1%; Nov: *m*=19.9%; *sd*=24.3%). Every interaction with the *Media Keys* required looking at the keyboard.

*Subjective Data*

The previously measured data could not be directly compared to data from the in-situ deployment. Particularly for privacy reasons we had not been able to measure anything related to the primary task during the in-situ deployment. In the peripheral task, errors and reaction time are not measureable because of the task's intrinsic motivation. But we designed the questionnaire in the lab study similar to the in-situ deployment to be able to compare subjective ratings.

All four devices were considered to be **easy to learn** (for all *User Groups* and *Interaction Styles median*=4 or higher[2]). The most **enjoyable** devices are *Graspable* (all: *median*=4 or higher) and *Touch* (all: *median*=4). *Media Keys* (all: *median*=3) and *Freehand* (all *median*=3 or lower) are moderately enjoyable. Ranking for **easy interaction** is similar: *Graspable* (all: *median*=5), *Touch* (all: *median*=4 or higher), *Media Keys* (In-Situ: *median*=5; Exp: *median*=3.5; Nov: *median*=4) and *Freehand* (all: *median*=3 or lower).

[2] higher and lower refer to a difference of at most one.

In terms of interaction in the periphery of the attention, **mental load** was low for *Graspable* (all: *median*=4 or higher) and *Touch* (all: *median*=4 or higher). During the in-situ deployment *Freehand* was considered to not be mentally demanding compared to the lab study (In-Situ: *median*=5; Exp: *median*=3; Nov: *median*=2.5) while *Media Keys* bothered users more during the in-situ deployment (In-Situ: *median*=2.5; Exp: *median*=3.5; Nov: *median*=4). **Distraction** was also low for *Graspable* (all: *median*=4 or higher) and *Touch* (all: *median*=3.5 or higher). For *Freehand* and *Media Keys* distraction was medium (Both: *median*=3). Similarly participants felt that they could interact **without looking at the device** with the *Graspable* (all: *median*=4 or higher) and *Touch* (all: *median*=3.5 or higher). *Freehand* was rated medium (all: *median*=3 or lower) followed by *Media Keys* (all: *median*=2.5 or lower).

Users stated that their **performance in the primary task** was not affected much while interacting with the *Graspable* (Exp: *median*=2; Nov: *median*=2.5) and *Touch* (Exp: *median*=2; Nov: *median*=3). *Freehand* (All: *median*=3.5) and *Media Keys* (Exp: *median*=4; Nov: *median*=3) however led to a bigger effect in the primary task.

## DISCUSSION

To assess the validity of our approach we compared the results from the in-situ deployment and the lab study.

### Comparison of Experienced and Novice Users

Most papers on peripheral interaction invoke the large amount of time it takes for peripheral devices to move to the periphery of attention, as reason for not using a lab setting. Of course integration into routines and everyday life cannot be achieved within the short duration of a lab setting. To assess whether previous acclimatization to the device has an effect on our results, we recruited novice and experienced subjects for our study, and did not find any significant difference. This implies, that in our case the results of the lab study were also valid for participants who had never used the peripheral device before (which is the common situation when conducting a lab experiment).

Although differences were not statistically significant, we observed slightly more errors in the primary task but also in the peripheral task (except for media keys) for experienced users. While our experienced user group is rather small and therefore suffers more from variance in the data, we assume that these tendencies are due to sloppier interaction with the peripheral device, which some participants had adopted throughout the in-situ deployment. However, this did not affect the overall comparison of interaction styles.

### Comparison of In-Situ Deployment and Lab Evaluation

By proposing a controlled lab experiment for peripheral interaction we do not intend to replace field studies or in-situ deployments. We rather propose to extend the spectrum of evaluation methods for peripheral interaction. To assess the validity of our approach and find out how our lab experiment fits into the design process we compared the results from the lab with the in-situ deployment.

| | In-Situ | Lab-Exp | Lab-Nov |
|---|---|---|---|
| Performance Primary Task | | ✓ | ✓ |
| Error Rate Primary Task | | ✓ | ✓ |
| Reaction Time Peripheral Task | | ✓ | ✓ |
| Errors Peripheral Task | | | |
| - Tracking issues for freehand | | ✓ | ✓ |
| (pause/play, volume) | ✓ | ✓ | ✓ |
| - Tracking issues for touch (pause/play) | ✓ | ✓ | ✓ |
| Gaze Ranking Peripheral Task | | ✓ | ✓ |
| Subjective Ranking | ✓ | ✓ | ✓ |
| - Learnability, Easiness, Enjoyment | ✓ | ✓ | ✓ |
| - Mental Load | ✓ | ✓ | ✓ |
| - Distraction | ✓ | ✓ | ✓ |
| - Gazing | ✓ | ✓ | ✓ |

**Table 1. Summary of findings. Grey check mark states that results differed (in case of median by more than one).**

Table 1 summarizes the results gathered in our experiments. During the in-situ deployment we could not measure performance, error rates or track where participants gazed, but we inquired about participants' experience in semi-structured interviews. We identified two technical limitations: errors in the tracking of the freehand gestures and misinterpretation of short gestures for touch interaction. Both problems might be responsible for the rather high number of errors in the peripheral task during the lab study. Furthermore, in this lab study, interaction with the graspable device showed the most promising results in terms of performance in the primary task and reaction time, errors and gazes in the peripheral task. Graspable interaction was also the preferred interaction style in our field study.

Only one result differed between the in-situ deployment and the lab – the rating for mental load, especially for freehand and media keys. Participants in the field were more bothered by the media keys than by freehand interaction. During the lab study both user groups rated mental load in a reversed way: higher for freehand and lower for media keys. We assume that in the lab participants were very focused on the two tasks, and therefore were very aware of the location of the media keys and all peripheral devices whenever a trigger asked them for their interaction. However, the tracking issues we detected for the freehand gesture recognition might have disrupted their interaction flow and caused higher mental load in the lab setting whereas in the in-situ deployment we did not ask for precise interaction with the secondary task (e.g., volume at a medium level) and therefore the tracking issues were not as problematic.

### Controlled Lab Evaluation vs. In-Situ Deployment

Looking at Table 1, it seems like we can learn more from the lab evaluation than from the in-situ deployment. This is hardly true but merely a question of perspective and goal.

In the lab we found usability issues, which would have been very helpful to know before handing the prototypes to the participants in our in-situ deployment and probably would have strengthened the results. Furthermore we were able to observe participants and analyze their behavior. This is especially interesting for visual attention. During the in-situ deployment we only had the input focus of iTunes as weak indicator of visual attention. In the lab we analyzed all gazes and found that participants glanced at the media keys for every interaction. In contrast, during the in-situ deployment media keys were the alternative with the least interactions with iTunes in focus. Checking if the application connected to the peripheral device is in focus is one way to assess if the interaction unfolds in the periphery, but it does not provide comprehensive insights on visual attention. However, we found subjective ratings to be in line (even with the in-situ deployment) with the number of gazes to the additional device from the lab study. This observation supports subjective ratings as a reliable measure. Generally the lab evaluation gave us coherent and consistent data on the subjective appreciation of the four different devices.

The results from the in-situ deployment lack rankings based on quantitative data in terms of performance and errors, as tracking this would have been a considerable invasion into our participants' privacy. However, from the in-situ deployment we learned about the integration of these devices in daily life. For instance we observed that the peripheral devices were successful but the media keys not so much. When participants had one of the peripheral devices, most interactions with the audio player were carried out through these devices instead of the mouse. When participants had the keyboard with media keys, they opted for the mouse instead. This could not have been predicted from the data collected in the lab where participants were asked to use the peripheral devices and the media keys. We further found that participants preferred to use the peripheral devices with their right hand (which was their dominant hand), although their left hand would be unoccupied when only interacting with the mouse. Additionally participants started to use the devices in a way that we did not anticipate (e.g., carrying around the touch device as it did not depend on a cable connection to the computer). To overcome the novelty effect, in-situ deployments for a longer period of time are absolutely necessary.

### Reflecting on the Lab Study Methodology

The comparison of experienced and novice participants as well as the results of the field and the lab study gave us coherent results (with the exception of the assessment of mental load as previously discussed). The measurements concerning the primary task – *performance* and *error rate* – can be used to assess the degree of disruption that is imposed by the peripheral task. *Reaction time* is influenced by the mental and visual preparation that is necessary to start the secondary task. Finally, *errors* in the peripheral task hint at usability issues with the peripheral devices. In summary, all results that we gathered in the lab did also show in the field. The difference is merely that in the field results were based on semi-structured interviews and thus subjective data while we could derive results from quantitative data in the lab. However, with coherent results – independent of them being based on quantitative or subjective data – the most important difference is the time needed to perform

the evaluation. Without having participants familiarize with the devices for several weeks we were able to detect many usability issues and establish a ranking of the devices in a potentially early design phase. Improving the devices based on this findings and then deploying them in-situ would most likely strengthen the results from the field and offer more detailed insights in real life daily usage, as the integration would suffer less from prototypical usability issues.

## CONCLUSION

We presented laboratory experimental methods to gather feedback on peripheral interaction systems in the first phases of the design process. We described the design of a controlled laboratory study composed of two tasks, a primary task requiring continuous interaction and a secondary task happening in the periphery. We showed that this method provided comparable results to field results, even with novice users who had never experienced our system.

In conclusion, our work aimed at enriching the number of evaluation methods available to designers of peripheral interaction. Indeed, almost all studies of peripheral interaction used field deployments to study the impact of specific system. These in-situ experiments enabled the study of the long-term appropriation of technologies "disappearing" from people's main focus. However to get to the "right" design, controlled lab experiments still provide invaluable benefits. They enable to gather more precise measures of attention, spot usability issues and more broadly compare alternative technologies. In the end, preliminary laboratory experiments such as the one describe here should enable more successful field deployments[3].

## REFERENCES


1. Bailey, B.P., Konstan, J.A. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. Computers in Human Behavior 22, 4, (2006), 685–708.
2. Bakker, S., van den Hoven, E., Eggen, B. Design for the periphery. Eurohaptics, (2010), 71–80.
3. Bakker, S., van den Hoven, E., Eggen, B., Overbeeke, K. Exploring peripheral interaction design for primary school teachers. TEI, (2012), 245–252.
4. Bakker, S., van den Hoven, E., Eggen, B. FireFlies : Supporting Primary School Teachers through Open-Ended Interaction Design. OzCHI, (2012).
5. Beaudouin-Lafon, M., Mackay, W. Prototyping Tools and Techniques. In Human Computer Interaction Handbook: Fundamentals. (2007).
6. Boff, K.R., Kaufman, L., Thomas, J.P. Handbook of perception and human performance. (1986).
7. Bogunovich, P. and Salvucci, D. The effects of time constraints on user behavior for deferrable interruptions. CHI, (2011), 3123–3126.
8. Dix, A. Evaluation Techniques. In Human-Computer Interaction. (2004).
9. Edge, D. and Blackwell, A.F. Peripheral tangible interaction by analytic design. TEI, (2009), 69–76.
10. Hausen, D., Boring, S., Lueling, C., Rodestock, S., Butz, A. StaTube: Facilitating state management in Instant Messaging Systems. TEI, (2012), 283–290.
11. Hausen, D., Richter, H., Hemme, A., Butz, A. Comparing Input Modalities for Peripheral Interaction: A Case Study on Peripheral Music Control. Interact, (2013), 162–179.
12. Hazlewood, W.R., Stolterman, E., Connelly, K. Issues in Evaluating Ambient Displays in the Wild : Two Case Studies. CHI, (2011), 877–886.
13. Hsieh, G. and Mankoff, J. A Comparison of Two Peripheral Displays for Monitoring Email: Measuring Usability, Awareness, and Distraction. Tech Report, (2003).
14. Kahneman, D. Attention and Effort. Prentice-Hall, 1973.
15. Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M. Heuristic Evaluation of Ambient Displays. CHI, (2003), 169–176.
16. Mark, G., Gudith, D., Klocke, U. The Cost of Interrupted Work : More Speed and Stress. CHI, (2008), 107–110.
17. Matthews, T. Designing and Evaluating Glanceable Peripheral Displays. PhD Thesis, (2007).
18. Matthews, T., Hsieh, G., and Mankoff, J. Evaluating Peripheral Displays. In Awareness systems: Advances in theory, methodology and design, (2009).
19. McCrickard, D.S., Chewar, C.M., Somervell, J.P., Ndiwalana, A. A model for notification systems evaluation - Assessing user goals for multitasking activity. CHI, (2003).
20. Olivera, F., García-Herranz, M., Haya, P.A., and Llinás, P. Do Not Disturb: Physical Interfaces for Parallel Peripheral Interactions. Interact, (2011), 479–486.
21. Pousman, Z. and Stasko, J. A Taxonomy of Ambient Information Systems: Four Patterns of Design. AVI, (2006).
22. Shen, X. An Evaluation Methodology for Ambient Displays. Journal of Engineering, Computing and Architecture 1, 2, (2007).
23. Stasko, J., McColgin, D., Miller, T., Plaue, C., and Pousman, Z. Evaluating the InfoCanvas Peripheral Awareness System : A Longitudinal, In Situ Study. Technical Report, (2005).
24. Visser, T., Vastenburg, M., Keyson, D. SnowGlobe : The Development of a Prototype Awareness System for Longitudinal Field Studies. DIS, (2010), 426–429.
25. Wensveen, S.A.G., Djajadiningrat, J.P., Overbeeke, C.J. Interaction frogger: a design framework to couple action and function through feedback and feedforward. *DIS*, (2004), 177–184.
26. Weiser, M., Brown, J.S. Designing Calm Technology. PowerGrid Journal, (1996).
27. Wickens, C.D., McCarley J.S. Applied Attention Theory. (2007).


[3] The event-based primary task is available for download and usage at: http://www.medien.ifi.lmu.de/team/doris.hausen/files/pi-evaluation/