TESY: A Usability Test-Driven Prototyping Assistant Connecting Designers with Crowd-Testers

FELIX KRETZER, human-centered systems lab, Karlsruhe Institute of Technology, Germany
ALEXANDER MAEDCHE, human-centered systems lab, Karlsruhe Institute of Technology, Germany

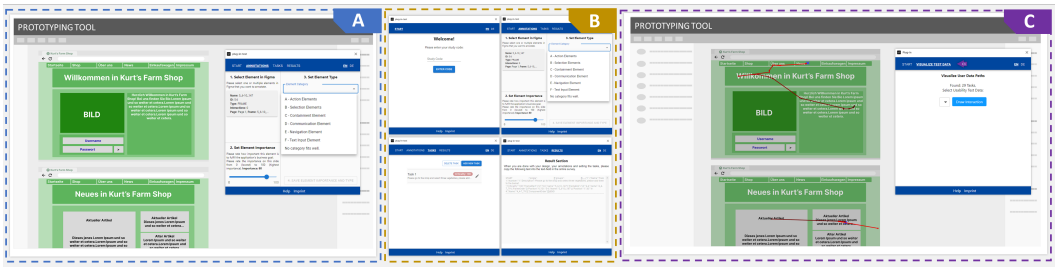

Fig. 1. Our plug-in *TESY* is directly integrated into prototyping tools (A) such as Figma. TESY allows in the first phase (B) to directly create asynchronous usability tests within the prototyping tool and link the task to annotated components. After the asynchronous usability tests are carried out, TESY later allows to display the usability test data within the prototyping tool linked to the design prototype's components (C).

In recent years, the availability of easy-to-use prototyping tools has empowered designers to create GUI designs. In parallel, a broad spectrum of usability testing tools have been proposed to support the collection of quantitative test data from crowd-testers. However, test specification and results are typically disconnected from created GUI designs and, therefore, difficult to translate into improvements. In this paper, we present TESY, a prototyping assistant integrating usability test specification and resulting test data. We implement TESY as a plug-in in the prototyping tool Figma. In a controlled lab study, we compare how 34 untrained designers create prototypes, specify usability tests, and improve the prototypes using the collected test data with and without TESY. We contribute by demonstrating how TESY empowers untrained designers to create enhanced GUI designs following a usability test-driven prototyping approach. Specifically, we demonstrate how TESY's capability of tightly integrating test data provided by crowd-testers into the prototyping tool leads to more data-driven and task-focused design improvements.

CCS Concepts: • **Human-centered computing** → *Usability testing*; *Empirical studies in HCI*; **User interface toolkits**.

Additional Key Words and Phrases: GUI Design; Usability Testing; Novice Designer

ACM Reference Format:
Felix Kretzer and Alexander Maedche. 2025. TESY: A Usability Test-Driven Prototyping Assistant Connecting Designers with Crowd-Testers. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW184 (April 2025), 24 pages. https://doi.org/10.1145/3711082

Authors' Contact Information: Felix Kretzer, felix.kretzer@kit.edu, human-centered systems lab, Karlsruhe Institute of Technology, Karlsruhe, Germany; Alexander Maedche, alexander.maedche@kit.edu, human-centered systems lab, Karlsruhe Institute of Technology, Karlsruhe, Germany.

ACM 2573-0142/2025/4-ARTCSCW184
https://doi.org/10.1145/3711082

1 Introduction

It has been known for years in practice and academia that the design of graphical user interfaces (GUIs) has a strong positive impact on the success of a software application (e.g., [30, p. 22]). In addition, it has been shown that the early integration of users into requirement analysis, design and prototyping leads to higher end-user satisfaction [40] and higher overall usefulness and usability [51, p. 477], [21]. Various methods have been proposed in research to enhance the design and creation of GUI prototypes in recent years. At the same time, a rich ecosystem of tools has emerged in practice to help experts create design prototypes (see, e.g., [3, 6, 18, 48]). However, a well-known problem in the area of GUI designs persists: *"designers tend to be brought in the process too late, to do superficial work on a system that has already been built"* [27, p. 10], resulting in significant design work being done by untrained (i.e., novice) users. In recent years, various approaches (e.g., [19, 26, 43]) have highlighted how experts and novices (e.g., [33]) can be supported in the creation of design prototypes. Plug-ins for existing prototyping tools, in particular, have proven promising.

With regards to including users in the design and prototyping phase, asynchronous usability testing following a crowd-testing paradigm has received particular attention in practice (see e.g., [37, 39, 50]) and research (see e.g., [5, 9]) in recent years. Here, the test facilitators (designers) and the crowd-testers are separated in time (i.e., usability tests are pre-scheduled) and space (i.e., usability tests are conducted without a direct connection between them) [23]. Asynchronous usability tests are less resource-demanding than their in-person alternative [5, 9]. While asynchronous usability test results are key for improving designs, they are typically decoupled from the prototyping tool. Consequently, recent reports in academia call for the *integration among tools* as "a key driver for tool adoption" [45]. Recently, first steps towards integration have been done. For example, Figma [18], one of the most widely used prototyping tools, has recently introduced the *developer mode*. In Figma, the editing screen can be switched from design to developer mode to integrate development tasks directly into the prototyping tool. Asynchronous usability testing tools, in practice, usually present the usability test data separated from the designs in the prototyping tools, leading to a disconnect between design activities and the presentation of results.

While there are first approaches proposed in practice to integrate usability test data into prototyping [49] and conceptual approaches in academia describing the integration of usability testing with prototyping tools [31] tailored to novice users, to the best of our knowledge, these approaches in academia have not yet been implemented and systematically evaluated with actual users. To address this research gap, we implement and evaluate TESY, a usability test-driven prototyping assistant visualizing user test data directly in Figma. Our work has three main contributions: First, we contribute by presenting *TESY*, an assistant that integrates usability test specification and resulting test data directly into the commercial prototyping tool Figma in the form of a plug-in. Second, we demonstrate how TESY empowers untrained designers in creating enhanced GUI designs following a usability test-driven prototyping approach. Third, we identify and describe differences in work patterns when test specification and resulting data is directly integrated into a prototyping tool.

2 Related Work

In the following we introduce conceptual foundations as well as related work to highlight relevant previous findings on usability testing, crowd-sourcing, crowd-testing, and prototyping assistance.

2.1 Usability Testing

Usability testing of designs has been covered extensively in academic literature for many years. The earliest literature can be found as early as the 1980s (e.g., [21]), and well-developed basics as early as the 1990s (e.g., [23, 42]). The need for usable GUIs has become more significant after the

introduction of personal computers into almost every area of today's life. Today, usability tests are supported by a variety of tools created for this purpose that focus on the different facets of usability testing [37, 39, 50]. In addition to the type of data collected (qualitative or quantitative), usability tests are usually differentiated according to whether the people carrying out the tests (*test facilitators* or *requesters*) and the participants in the test (*test participants*) are separated in space or time. If there is a spatial separation, the term *remote usability testing* is used. Castillo et al. [12] describe remote usability tests as generally cheaper compared to *in-person usability* [23, p. 229] when, in addition to a spatial separation, there is also a temporal separation. In the case of spatial separation, it is possible to determine ex-ante how the system will behave for given inputs. However, the test facilitator can no longer adjust the reaction patterns during the tests. Due to the lack of time restrictions, asynchronous usability tests are often rated more favorably in studies (e.g., [5, 9]). Existing research shows that more usability problems are identified when creating predefined tasks. Bruun et al. [10] compared how successfully test participants identified usability problems when they followed defined tasks or freely explored prototypes. When creating the tasks, it can be helpful to reflect which tasks are *critical*, and *relevant* [2, p. 67] for the software. In addition, it can help the test participants reflect on the application's business goal during usability tests [28].

2.2 Crowd-Sourcing and Crowd-Testing

Today, crowd-sourcing is being applied in different areas of the software development process. Crowd-testing is a contemporary approach in software development that outsources different testing activities to crowd-workers.

Crowd-testing the usability of GUIs has shown to be a scalable way to find test users and collect quantitative and qualitative test data (e.g., [24], [25]). There are numerous examples of research projects incorporating crowd-testing, such as, e.g., *CrowdUI* [44], a tool that lets crowd-workers arrange UI elements on to-be-tested web-designs, peer-evaluate modifications of other crowd-workers, and gather survey-based feedback on web-designs. *Apparition* [32], allows requesters to utilize crowd-workers to improve GUI prototypes based on sketches and natural language descriptions of desired functionality. *SketchExpress* [35] solves a similar challenge as *Apparition*, but focuses on interaction behaviours. Here, requesters share initial drawings of a GUI prototype with verbally described interaction behaviors with crowd-workers that then create the interactions for the GUI prototype. *CrowdStudy* [41] demonstrates a framework for crowd-sourced automated usability testing of websites with results mainly presented as metrics or visualizations such as heat maps. In contrast to our approach, *CrowdStudy* focuses on the website instead of GUI prototypes and does not integrate test data into the development environment. *Voyant* [52], breaks down tasks to collect perceived feedback on design (such as posters or graphics) into sub-tasks that are sent to *Amazon.com's Mechanical Turk*. While some feedback is linked to single elements (here, what element of the design was noticed first), no usability test data beyond first impressions is directly linked to components within the design environment.

Crowd-testing usability of websites or GUIs has been frequently compared to traditional lab settings, e.g., Liu et al. [36] find advantages and disadvantages for both methods, but state that "major [usability] problems [..] were identified by both lab test participants and crowd-workers" [36, p. 7] in their study. Crowd-testing is also often evaluated in comparison to experts' feedback. Schneider et al. (2013) [47] compare website usability tests conducted by crowd-workers with expert data. They conclude that crowd-workers can be recruited in large numbers, at low cost and can be able to find the most pressing usability problems. Additionally, *CrowdCrit* [38] shows that crowd-sourced feedback can approach expert design evaluations.

Further research has looked into comparative usability testing and benchmarking, such as Deka et al. [17] that present ZIPT, a tool enabling feedback requesters to perform task-based usability

tests on third-party apps using crowd-workers. The usability test data is present on dashboards as individual traces (clicks marked on screenshots), performance metrics (completion rate, time on task, and number of interactions), qualitative feedback (based on questions predefined by requesters), and flow visualization [17, p. 729]. Additionally, research has been conducted on improving crowd-sourced usability test coverage, such as, e.g., Chen et al. [13] (reducing redundant paths in usability tests by introducing *interactive event-flow graphs* and *GUI-level guidance* for user interface testing), and real-time collaboration in crowd-sourcing, such as work by Lee et al. [34], that reuses *Apparition* to evaluate how requesters communicate and collaborate with crowd-workers.

Multiple research projects, most relevant to the approach evaluated in this paper, provide an in-situ integration of crowd-sourcing within dedicated design or development tools from practice. However, none of these projects evaluate the context of GUI prototyping.

In the area of creative writing support, both *Soylent* [7], and *Heteroglossia* [29] incorporate crowd-sourcing directly into dedicated writing tools (here MS Word and Google Docs). *Soylent* [7] is a writing support tool that sends parts of a text on-demand to crowd-workers that shorten, proofread, or edit text written by a requester. Notably, Soylent integrates into MS Word to specify the tasks, control the crowd-work, and present results within MS Word, which can be incorporated directly into the current writing. Huang et al. [29], present *Heteroglossia*, a Google Docs add-on that provides creative writing support by sharing snippets with crowd-workers that then generate *follow-up story ideas*. Remarkably, the follow-up story ideas are integrated as comments in Google Docs, offering direct tool integration. To support brainstorming, Andolina et al. [4] present *Crowdboard*, a real-time ideation support system augmenting brainstorming sessions on a (digital) whiteboard with real-time ideas from crowd-workers. Crowd-workers' feedback is presented in real-time and in situ as they directly attach comments to ideas while feedback requesters work on the whiteboard.

Palani et al. [45] find that creative practitioners value *integration with current workflows* as the second most crucial aspect for creativity support tool adoption. Within the category *integration with current workflows*, their "survey shows practitioners valued integration across tools the most" [45, p. 5]. While multiple approaches integrate crowd-sourcing-based processes and results in situ into some dedicated development environments (such as MS Word or Google Docs), comparable approaches are currently missing in the area of usability testing of GUI prototypes.

2.3 Prototyping Assistance for Novices

In literature various assistance systems have been proposed, often as plug-ins, to aid (novice) designers with prototyping tasks, enabling direct interaction with the prototyping tool and design prototypes. Lee et al. proposed a categorization of prototyping assistance systems, distinguishing systems that help find the suitable samples, help import existing components, and systems that automate design tasks [33]. Lee et al. themselves introduce a browser-based plug-in, *GuiComp*, that provides real-time feedback: an attention map, suggestions for similar GUIs from RICO [16], and seven metrics (e.g., element size, balance, alignment, font size). Lee et al. design for and evaluate GuiComp with novice users. Hereby, they address that it is not uncommon for users with little to no experience to take over the creation of GUI prototypes in practice [27]. Hegemann et al. [26] introduce CoColor, a technique and plug-in to help with color exploration, assignment, and refinement. Bertram et al. [8] present a plug-in that enables automated usability evaluation of UI prototypes based on metrics. Recently, Kretzer and Maedche [31] proposed a quantitative test-driving prototyping approach that suggests integrating usability tests into prototyping tools. Specifically, they articulate three generic design goals: First to integrate the annotation of components and the usability testing task specification within the prototyping environment, second to support quantitative usability testing and third to present the test data specific, justified, actionable and connected to the components within prototyping environment [31, p. 2-3].

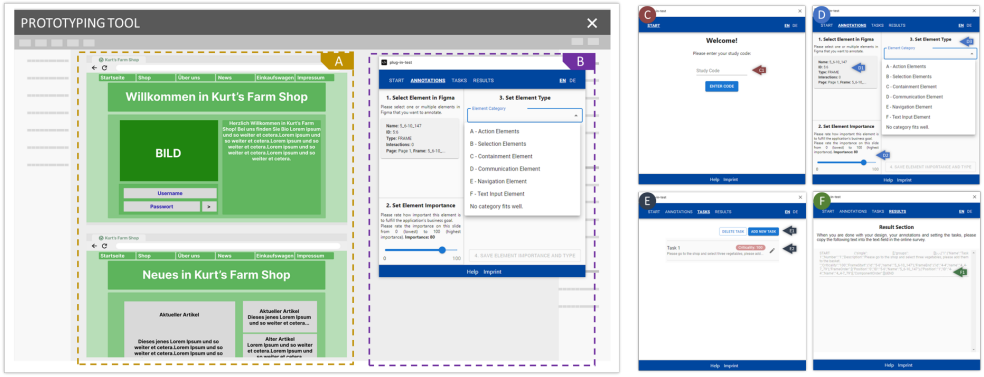

Fig. 2. Our plug-in TESY can be integrated into prototyping tools. Area A shows a design prototype within the prototyping tool. TESY (B) interacts with the design prototype. Study participants must enter their ID (C) before proceeding to the component annotation (D), usability task specification (E), and data export page (F).

3 System Design

Our design features were motivated by the design goals identified by Kretzer and Maedche [31, p. 2-3] for integrating usability test specification and data into dedicated prototyping tools, namely:

"**Design Goal 1**: *Assist designers in annotating components (DG1-A) and defining tasks for usability testing (DG1-B) tightly integrated in the prototyping tool (DG1-C).*"
"**Design Goal 2**: *Support designers in collecting quantitative usability test data asynchronously for the prototype based on the defend tasks.*"
"**Design Goal 3**: *Present the collected data from usability tests in the prototyping tool (DG3-A) directly connected to the annotated components (DG3-B) in a specific, justified and actionable form (DG3-C).*"

While the referenced work did a formative evaluation of the design goals, so far, no summative work has evaluated a system built on the design goals. While there are already summative evaluations of tools in some domains (e.g., creative writing support) that integrate the creation of tasks and the presentation of the results directly into editors (e.g., in MS Word or Google Docs), there is no corresponding work for the creation of GUI prototypes that integrates and evaluates both the creation of usability tests and the results in practical editors for the often neglected group of novice users, although in practice they often take on design tasks.

For our summative evaluation in this paper, we implemented a plug-in for the prototyping tool Figma [18] that supports annotation, usability task specification, and visualizes usability test data within the prototyping tool. We additionally implemented a web-based interface to collect asynchronous usability test data and a web-based representation of usability test data that serves as a benchmark to evaluate our plug-in. In the following, we present our system and its features. Our system supports all three phases: 1) Creating the initial design prototypes (supported by TESY). 2) Collecting usability test data on the design prototypes (supported by a web-based usability test prototype). 3) Improving the design prototypes based on the collected usability test data (supported by TESY and as a baseline the web-based visualization).

3.1 Connecting Prototyping and Testing through Annotations and Task Specification

We designed TESY to support annotating components (such as, e.g., buttons or text elements) (DG1-A) and creating usability test tasks within the prototyping tool (DG1-B). Figure 2 shows a screenshot of our plug-in TESY integrated into a schematic prototyping tool (in our case Figma).

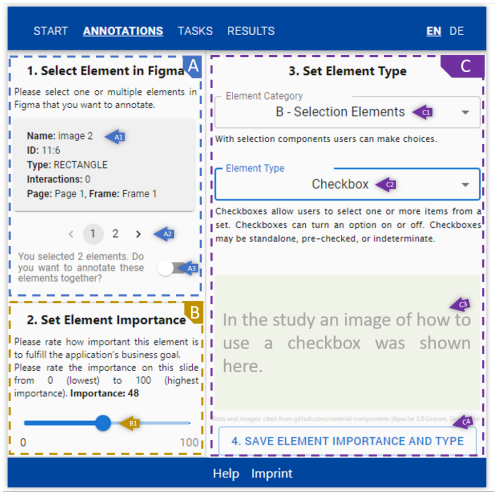

Fig. 3. Annotation Interface: Area A displays the components selected in the prototyping tool for annotation. Area B allows reflection on the importance of each selected components by annotating the importance. In area C, components are matched with an element category and element type from *Google's Material Design* (explanatory texts quoted from [20]).

TESY can directly access information about the design prototype by accessing the Figma Plug-In API, thereby gaining read/write access to the document structure of an GUI prototype in Figma. This deep integration is necessary to first, map components in Figma to the annotation and the solution paths for usability test tasks, and second, to overlay usability test data within Figma. TESY's position can be adjusted, and the entire plug-in can be closed without losing entered data (a design feature identified in a previous formative study [31]). TESY is organized in different pages: On the first page (C in Figure 2), participants can enter their study ID (C1) to match the collected data with designs later (a technical requirement in the study). The annotation page (D in Figure 2) allows users to select one or multiple components in the prototyping tool, verify their selection (D1), assess the component's importance for the application's business goal (D2), and select a category for the component (D3). The third page (E in Figure 2) allows the setup of usability testing tasks (E1, E2). On the last screen (F in Figure 2), TESY shows all entered data (F1) as a JSON for exporting (also a technical requirement to export data in the study).

Annotation. Based on design goal DG1-A, we implemented the option to annotate design components to let users reflect on their design and potentially use the annotations for analyzing the test data. Figure 3 shows how TESY enables annotating components directly within the prototyping tool. Users can select one or multiple components within the prototyping tool by clicking on them and see their selection visualized in the plug-in (A in Figure 3). The selection is shown, and users can check if the correct label of their desired selection are displayed (A1). If users want to select multiple elements, they can click through the selected element (A2) and decide whether to annotate multiple elements with the same information or annotate the selection as one group (A3). Users also set their selection's importance (B in Figure 3) within TESY, to let users reflect on which components deserve the focus of the user's attention [31]. Ultimately, users assign an element type (C in Figure 3) to their selection (e.g., button, text box). TESY helps users find the right element type by proposing element categories (C1) and element types (C2), supporting every selection with explanatory texts, and showing images of the element types as an example (C3). TESY proposing

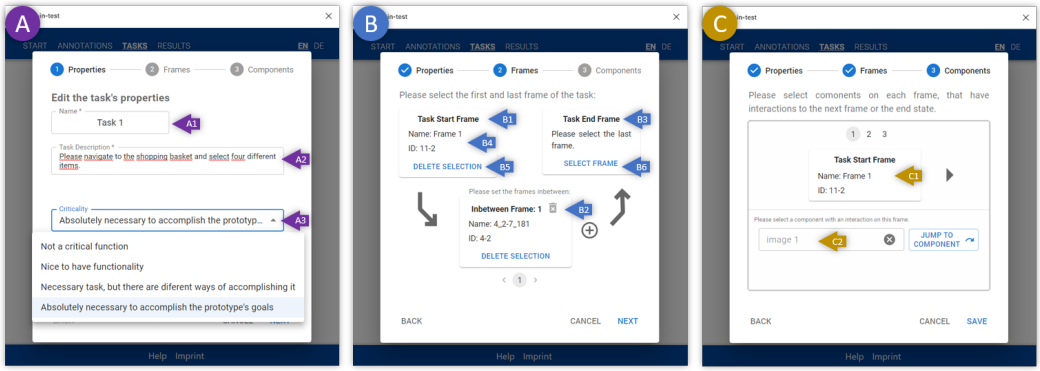

Fig. 4. Interface to create usability task specifications. Area A allows users to add new tasks, set up the task's name and instruction, and rate the task's criticality for the application's business goal. Area B lets users define the task's start, end, and intermediate frames. In area C, users can select components on each defined frame that own an interaction, making the task clickable.

element categories was a design choice directly motivated by DG1-A. Hereby, we wanted to assist novice designers by presenting a predefined list of possible element categories instead of using text boxes. To follow industry practice, we used the component structure and component's explanation texts from *Google's Material Design* [20].

Usability Test Specification. To enable participants to later improve their design prototypes based on test data, TESY offers the function to specify usability tests (following DG1-B). Figure 4 shows how the usability test can be specified directly within the prototyping tool. Users can define metadata on their usability tasks (A in Figure 4). There, the task name (A1), the task instructions (A2), and the criticality of the tasks (A3) can be set. We guide users to rate the criticality of the tasks [31]. Thereby, they can prioritize the tasks (also during test data analysis) and reflect on the importance of each task. On the next page (B in Figure 4), users can set the first (B1), last (B3), and all intermediate frames (B2) to solve a usability test task. Usually, a task starts and ends on predefined frames - though start and end frame can be the same. To add a frame to the task, a designer clicks on *select frame* (B6) and then clicks on a frame in the prototyping tool. The selected frame's name (B4) is displayed to allow for controlling the selection. Selections can also be deleted (B2). At the end of task specification (C in Figure 4), TESY allows selecting components on the frames (C1) that own an interaction (C2). The interactions (e.g., a click on a button) must allow a transition between the frames so that users can navigate from the start frame of a task to the end frame. Frame and component selections in the prototyping tool are read by TESY through the prototyping tool API (e.g., the Figma Plug-in API). Using the predefined annotations, components that have interactions are suggested. The features for setting up the usability test tasks are motivated by the first design goal (DG1-B). While similar work [17] did not integrate test task creation and test data presentation into prototyping tools, we have created the test task specification close to other approaches from literature, such as, e.g., ZIPT [17], where usability tests are described by a task text, and can include a starting point (frame or element of an design).

3.2 Supporting the Collection of Usability Test Data

For the collection of usability test data, in principle, any commercial tool can be used that allows access to the raw data and can import the design prototypes previously created in the prototyping tool. To collect test data in this study, we implemented a temporary prototypical usability testing

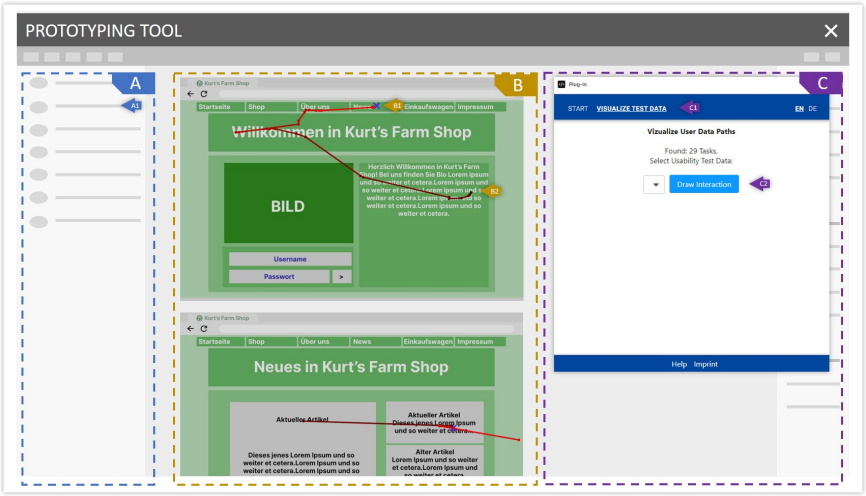

Fig. 5. TESY (area C) supporting the display of usability test data as overlay (area B) in prototyping tools. With TESY, users can select and draw single usability test sessions as overlays over the design prototypes in the prototyping tool. The overlays are added as frames that can be hidden (area A).

tool, as a website, by ourselves motivated by the second design goal. We envision that the test tasks will be directly available to the crowd-workers on the temporary test website after describing the usability test tasks. Users can then use existing crowd-working platforms and direct participants to the test website. Our test website displayed the task next to the respective design prototypes. Mouse movements and clicks were logged during an interaction with the design prototype. Finally, the testers could articulate whether they had solved the task. The collected data is necessary to understand later how testers approach the task. We again build this solution close to other approaches from literature, such as, e.g., ZIPT [17], where crowd-testers' interactions are recorded, and crowd-workers are required to self-report whether they believe they have solved the task correctly.

3.3 Connecting Usability Test Data with Prototyping

Our study aimed to find out how users can be better supported in interpreting usability test data within the prototyping tool, as part of the third design goal. A exemplary representation of TESY's test data integration into the prototyping tool is shown in Figure 5. In order to provide a baseline, we also developed an independent web-based interface shown in Figure 6. This web-based interface refers to the current landscape of usability testing tools from practice (see, e.g., [37, 39, 50]), but also to not-integrated test data representations in literature (e.g., [17]). We purposely decided to allow for the same visualizations in both TESY and the web-based usability test data interface, since we only wanted to evaluate the effects of integrating test data in situ into prototyping tools.

Integration of Usability Test Data into the Prototyping Tool. TESY allows users to select individual usability test sessions (area C in Figure 5) and draw the interactions as an overlay in the prototyping tool (area B in Figure 5). The overlays are displayed in the hierarchy of the prototyping tool (A in Figure 5) and can be hidden (A1). Overlays are drawn in the prototyping tool via the prototyping tool's plug-in API. The overlays show clicks as blue crosses (B1) and mouse movements as red lines (paths start with a slightly darker red). Users can display multiple user testing session as overlay over a single frame and thereby compare data from multiple usability tests.

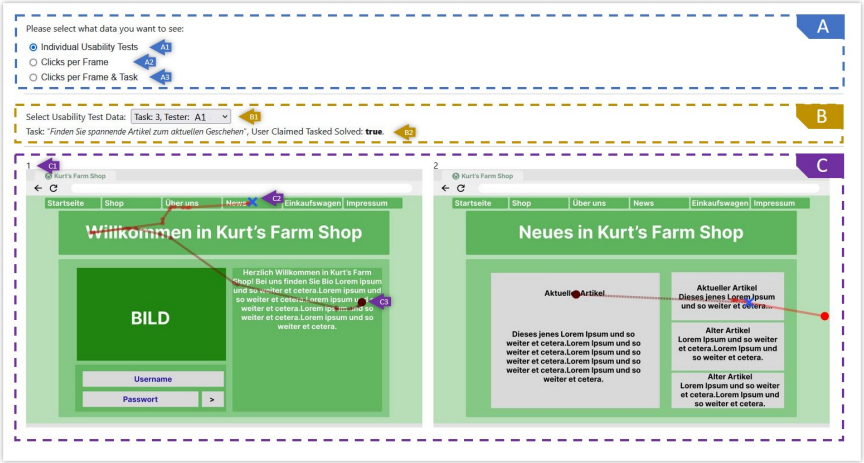

Fig. 6. Web-based interface of usability test sessions. With the data display mode (in area A) users can select data from a single tester filtered by tasks, clicks from all tests filtered by tasks, or clicks from all tests for this design prototype. The filter can by applied in area B. Area C visualizes the usability test data (blue crosses symbolize clicks, red lines mouse movements), shown as overlay over the selected design prototypes.

Web-based Usability Test Data Interface. Participants can select the data display mode on the first quarter of the page (A in Figure 6). They can either select to display single usability test data sessions (A1), an overlay of all clicks by all participants for each design prototype frame (A2), or filtered by task (A3). In the second area (B in Figure 6), participants can select individual testing sessions to be displayed (B1). Depending on the selection, usability testing data is shown as an overlay (C in Figure 6) of the design prototype frame (C1). Clicks are displayed as blue crosses (C2), a mouse movement as red line. The start of each mouse movement is marked by a darker red (C3).

4 Evaluation Study

We conducted a mixed method between-subjects study with multiple phases to find out how TESY affects novices during initial designs and later during design improvement based on usability test data. A particular goal was to find out how the quality of the designs created is affected by TESY and how the working process of the participants in our study changes. The first three phases of the study were conducted in our university lab, the last phase was conducted using crowd-workers with experience in usability design and usability testing. The study design was approved by university's ethics committee and the data protection officer in advance.

4.1 Procedure and Tasks

Our study (study design shown in Figure 7) is organized in four main phases: In the **first phase**, after receiving a basic introduction on prototyping and the tool Figma, the invited 20 participants were given a principal design task and implemented this task using the prototyping tool Figma including our TESY plug-in. Besides designing the prototype, the participants had to annotate components with the plug-in and create tasks for usability testing. Furthermore, as a baseline 14 participants worked on the same design task without the TESY plug-in as a control group. In the subsequent **second phase**, usability test data was collected. We leveraged the 14 baseline participants and hired 14 additional participants to perform usability tests as specified with TESY. In the **third phase**, participants from the first phase improved their designs based on the collected

usability test data. They were randomly divided into two groups and given access to the test data via the plug-in or decoupled via a separate website. Finally, in the **fourth phase**, we presented the prototyping results to experts in usability design and usability testing hired on the crowdsourcing platform Prolific [46] and asked them to rate the design prototypes and created tasks.

Principal Design Task: We wanted to give participants a design task that each participant could empathize with and have a solution at its core that participants might have interacted with in a similar way before. The scenario of the main design task introduced the imaginary farmer Kurt, who runs a farm in the immediate vicinity. In order to sell products directly to customers, he would like to offer his own web store where customers can flexibly compile their own organic boxes and order them directly. In the scenario different requirements were presented (overview of all products, recognizable regional focus, an about-us section, a news section and a shopping cart). Participants were asked to implement only a selection of the requirements, but to create clickable prototypes so that the usability tasks could be performed on an interactive version of the prototype. In addition, participants were given other fixed rules and recommendations. For example, participants were not allowed to insert images from the web into their designs, and were only allowed to work with placeholders for the images, as otherwise too much time would have to be spent searching for usable images.

4.2 Participants

Overall, we recruited four different groups of participants for our study. No participant took part in more than one group. *Group 1*, *Group 2* and *Group 3* were recruited from a university student panel, participants for *Group 4* were recruited through *Prolific*. In all groups, more potential participants were invited than had actually taken part (participants were sorted out, for example, if they had a technical defect or did not meet the requirements for participation). The groups formed from all valid participants were composed as follows:

Group 1 consisted of 20 students (13 male, 7 female) recruited from a university panel, that received the task to design initial prototypes in a first session, and improve the prototype in a second session a few days later using different treatments. Participants in this group had an average age of 23.7 years (σ = 3.57), currently students, and have been studying for on average 4.10 years (σ = 2.70). In this group, most participants completed an high school diploma and about half completed a bachelor's degree. 19 participants reported no prior experience and 1 participant reported 1 year or less prior experience with the creation of visual design. Similarly, all participants reported no experience judging visual designs. No participant failed any of our attention checks, and thus all 20 students were admitted to the study.

Group 2 consisted of 14 students (7 male, 7 female) recruited from the same university panel. In this group we asked participants to design initial prototypes without any help (as a control to *Group 1*). Subsequently, all participants in this group were asked to perform asynchronous usability tests on the designs of *Group 1*. Participants in this group were on average 25.5 years old (σ = 3.17), have been studying for on average 4.40 years (σ = 2.27). 11 students reported no prior experience with the creation of visual design and 12 students reported no experience judging visual designs. 3 and respectively 2 students reported 1 year or less of prior experience designing and respectively judging visual designs. No participant failed any of our attention checks, and thus all 14 students were admitted to the study.

Group 3 comprised 14 students (9 male and 5 female) who were recruited from the same university panel. In this group, we asked participants to conduct asynchronous usability tests on the designs created by Group 1. The participants in this group had an average age of 24.5 years (σ = 4.33) and had been studying for an average of 3.64 years (σ = 1.64). 13 students reported no prior experience with creating visual designs, nor experience in evaluating visual designs.

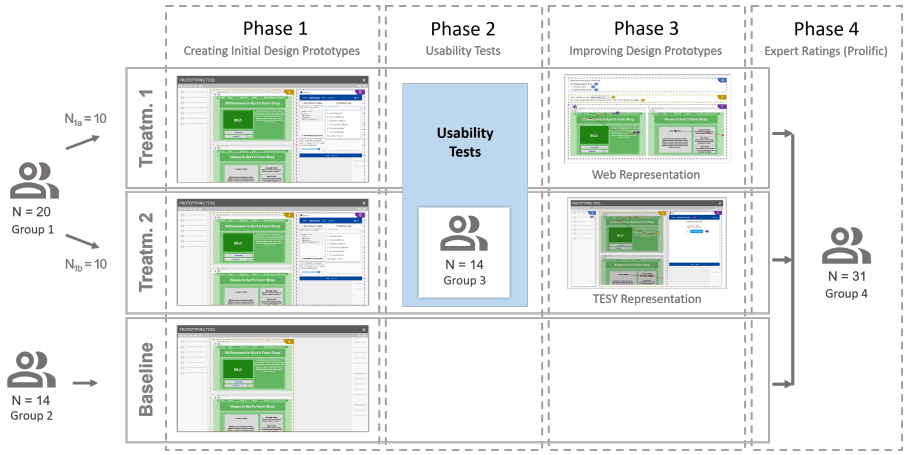

Fig. 7. The study comprised four phases and involved four groups of participants across three configurations: *Baseline*, *Treatment 1*, and *Treatment 2*. *Group 1* participants were randomly allocated to either *Treatment 1* (web-based presentation of usability test data), or *Treatment 2* (integrated usability test data with *TESY*). In *Phase 1* participants created initial designs, followed by usability testing in *Phase 2*. Participants refined their designs in *Phase 3* using usability test data. In *Phase 4*, experts evaluated all prototypes (initial and revised).

Group 4 consisted of 31 experts (25 male, 6 female) recruited on the crowd-working platform *Prolific* that evaluated the designs of participants in the two previous groups. On Prolific, we chose to only invited participants located in Germany and fluent in English and German, as our design prototypes from previous study phases often contained German texts and additionally German tasks. Furthermore, we only invited participants older than 18 years, that stated knowledge of *UX, A/B testing, or UI design*. While we originally invited 40 participants, we chose to exclude 9 participants due to failing our manipulation checks (2 participants exclusively failed our attention checks, 5 participants failed our comprehension checks and 2 participants failed both checks). Participants in this group were on average 30.9 years old (σ = 6.8). Participants had on average 2.9 (σ = 4.3) years of experience in creating visual design, and 2.4 (σ = 3.8) years of experience in evaluation visual design. While all 40 participants on Prolific were paid, only data-sets from the remaining 31 participants were considered.

First Phase - Initial Prototype Creation: Participants in this part of the study joined the study on-site in a lab setting. The study lasted 120 minutes and began with the participants agreeing to our informed consent form. Then participants started to fill out an online questionnaire asking for demographic data and experiences with creating and evaluating graphical user interfaces. Next, the participants watched a brief video explaining common methods of the design tool. After the video, the participants received a brief pre-scripted in-person guided tour explaining the scenario, principal design task, general rules, design tool and our plug-in TESY including the tasks related to the plug-in. The information provided in the guided tour was also given to the participants as a written handout. At this point - usually 30 minutes passed - the participants were asked to start working on the principal design task. We instructed the participants to work on the principle design task for 50 minutes and spend 20 minutes annotating components and preparing usability tests with TESY. While we gave them a fixed time-budget for both tasks, we explained that design, annotations and setting up tasks for usability testing can be done in parallel or any other order as long as participants did not exceed the time-budgets for each individual part. After designing and using TESY, participants returned to the questionnaire for the last 10 minutes, where they

entered data from TESY, were asked for their subjective experiences with the TESY in the form of free text-fields but also in the form of task-load, usability and creative related questions (following Hegemann et al. [26], and asking for SUS [1], CSI [11] and NASA Raw TLX [22]), since we wanted to control that our plug-in did not significantly affect task load, creativity or the overall usability with the design tool negatively. Participants in this part of the study were paid a fixed amount of 27 euros for participating in the study. Additionally, participants could achieve two extra payments of 5 euros if the usability of their designs and relevance of all valid tasks was rated among the best 30% of each group.

Second Phase - Usability Test Data Collection: In this phase we collected usability test data. We leveraged the baseline participants to perform usability tests as specified through TESY. In principle, this phase is independent of TESY in the sense that any asynchronous usability test tool can be leverage to collect usability test data on the basis of the prototype and the defined tasks.

Third Phase - Prototype Improvement with Test Data: In our third phase, the participants from the first phase returned to improve their designs. Since our participants learned the usage of the design tool in the first phase, they were not required to watch the explanatory video again. They were again given an online questionnaire including demographic data and prior experience as well as a brief pre-defined in-person guided tour describing the general rules and further details about the task at this stage. Participants were tasked to improve their previous designs based on the collected usability test data. Half of the participants were provided with the test data via the web-based representation we provided as described above. The other half were provided with the collected test data directly integrated into the prototyping tool through TESY. The overall goal of the improvement phase was to improve the design so that errors that might have occurred with the test users would no longer occur in the future. Participants had 15 minutes to read in, start the questionnaire, and a brief pre-scripted in-person guided tour. There were 45 minutes for improving the designs. Finally, participants answered similar questions about the usability, task load, and creativity of the system they used and listed the improvements on their design. In total, this part of the study took 75 minutes. Participants received 17 euros and again had the option to receive 5 euros as a bonus payment if the improvements made were rated in the top 30 percent of all improvements per treatment.

Fourth Phase - Prototype Evaluation by Experts: In order to evaluate the created design prototypes and usability tasks, we invited experts from the crowd-sourcing platform Prolific. For this purpose, participants received a questionnaire using Prolific for sourcing participants. After consenting to the privacy terms and conditions of participation, the questionnaire asked for demographic data and data on previous experience with visual design and the evaluation of visual design. Participants were then introduced to the principle design task, and the rules under which previous participants had created the designs (e.g., not including images from the web) were explained to them. Then, participants received three randomly selected design prototypes created without assistance (baseline). Participants rated the design prototype's goal fulfillment, usability, and interactability. Subsequently, the participants received nine randomly selected design prototypes created with TESY or the web representation of the usability test data. The participants were either shown the initial design prototypes or the design prototypes after improvement. We ensured that no participant received an initial design and later an improved design of the same prototype to not bias the participants. Additionally, the created usability tasks were shown. The designs were presented as images in the questionnaire. The interaction areas and interaction paths were visualized for the participants. Study participants received 6.57 euros for their participation, with a predicted 30 minutes of processing time (payment of 13 euros per hour, analog with the other study phases). The average time to complete the phase of the study turned out to be 25.47 minutes.

4.3 Data Collection and Analysis

Quantitative Data Collection. In the different phases of our study, we collected quantitative data in several ways. In Phase 1 and Phase 3, we chose to collect subjective survey-based data in the form of the *Creativity Support Index* (CSI, [11]), System Usabiltiy Scale (SUS, [1]) and unweighted NASA Task Load Index (*NASA-TLX* Raw, [22]), comparable to related work in our field (see e.g., [26]). We always assessed the overall system (design tool and our plug-ins or other treatments), in order to test whether our system introduced significant downsides in regards to task load or overall usability. In additional, in our third phase we asked participants to express their agreement on two statements using five point Likert scale. The statement were, first: *I understood what problems the users had during the usability test.*, and second: *I feel very confident that I could significantly improve the design with the data presented.*

In order to be able to examine how participants actually worked with the prototyping tool and our TESY assistant, we also made screen recordings of all participants. Furthermore, eye-tracking data using Tobii Pro Nano devices was collected. Eye-tracking data was collected from six randomly selected participants in each of the Phases 1, and 3.

The crowd-workers from Prolific rated design prototypes created without TESY in regard to goal fulfillment, usability, and interactability on a 9-point Likert scale. Additionally, for design prototypes created with TESY, the crowd-workers rated the usability testing tasks on a 7-point Likert scale (comprehension, feasibility, relevance). Afterward, the crowd-workers were asked to rate the initial and improved designs on a 9-point Likert scale. The crowd-worker were asked to rate how both versions fulfilled the principle design task, how the usability was estimated in each case, and whether the usability test tasks could be solved with the presented design prototypes.

Qualitative Data Collection. In the first phase, participants were asked in an open text field what they liked and disliked about using TESY. In the second phase, participants were asked what they liked or disliked about the task. In the third phase, depending on the treatment, participants were also asked either what they liked or disliked about the general task or the task using the plug-in. In the third phase, we also asked about the specific improvements the participants had implemented based on the usability test data and where the participants had problems understanding the data.

5 Results

To examine the effects of TESY on the created tasks, the usability of the initial design prototypes, and the improved design prototypes, we analyzed the data collected from the participants. In addition, we examined the quantitative results of the experts who evaluated the respective tasks and usability. Finally, we analyzed the collected eye-tracking data and screen recordings. We compared them to the collected qualitative data to better understand how TESY impacted the work patterns in prototyping.

5.1 Quantitative Results

Our participants had little prior experience in creating and evaluating visual designs in the form of prototypes. We collected the NASA Raw TLX [22], CSI [11], and SUS [1] for all treatments to ensure that our treatments did not significantly negatively impact the overall system used (consisting of either the design tool only, the design tool with the plug-in, or the design tool with the website on which test data were presented). In each case, we tested for normal distribution using the Shapiro-Wilk test and then for equal variances (although the study design already suggested equal variances). Depending on the outcome of the initial tests, we performed either *signed-rank Wilcoxon tests*, *Student's t-tests*, or *Welch's t-tests*. Table 1 shows mean values and standard deviations of both conditions in Phase 1 and Phase 3.

Table 1. NASA Raw TLX, SUS and CSI during first and third phase.

Parameter	First Phase				Third Phase			
	No Plug-In		TESY		Website		TESY	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NASA Raw TLX	**51.31**	**11.86**	**53.00**	**10.83**	**42.83**	**9.95**	**39.50**	**12.64**
SUS	**65.89**	**14.73**	**56.00**	**15.72**	**63.50**	**15.82**	**60.25**	**18.80**
CSI	**67.10**	**18.71**	**57.77**	**16.04**	**73.53**	**13.11**	**61.80**	**22.12**

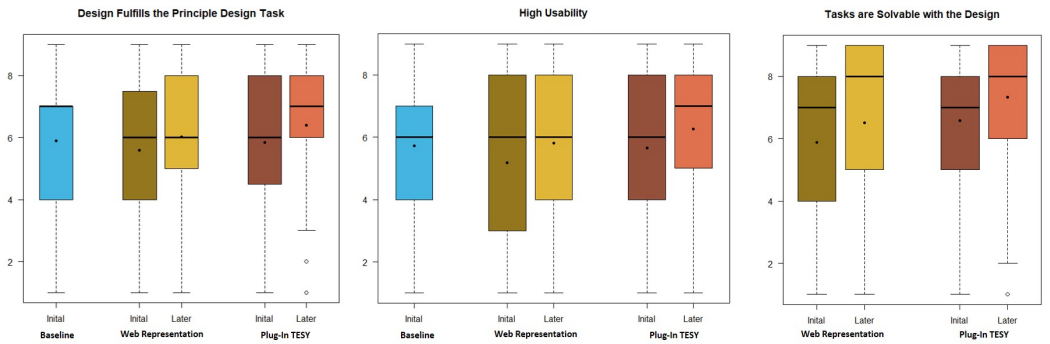

Fig. 8. Boxplots of experts' ratings of the design prototypes in each phase. Experts rated their agreement with three statements *the design prototype perfectly fulfills the principle design task*, *the design prototype has a high usability*, *the usability tasks can be solved with the design prototype* on 9-point Likert scales. Experts rated design prototypes that were not improved and did not use our Plug-In (Baseline), design prototypes using the the web-based representation of the usability results for improvements, and design prototypes using TESY for visualizing usability test data for improvement. Black dots represent means.

We find that TESY had no significant negative influence during the first phase ($n_1 = 20, n_2 = 14$) on task load, creativity support or overall system usability. As results of the Shapiro-Wilk test we performed *Student's t-tests* and found no significant negative influence for NASA RAW TLX ($p = 0.8337$), CSI ($p = 0.2501$), or SUS ($p = 0.1465$). Also, we find that TESY had no significant negative influence during the third phase ($n_{1,2} = 10$) on task load, creativity support or overall system usability: NASA RAW TLX ($p = 0.5203$), CSI ($p = 0.16618$), or SUS ($p = 0.6807$).

In our third phase, we asked participants to express their confidence in their improvements they have made and whether they understood what problem the test users encountered. Following de Winter et al. [15], we used a *Wilcoxon rank sum test*. While using the web representation of the data and the integration into the prototyping tool did not lead to significant different means, our participants using the plug-in reported slightly higher agreement to the statement *I feel very confident that I could significantly improve the design with the data presented* (means of 3.5 vs 3.0 on a 5-point Likert scale, $n_{1,2} = 10$).

In the final study phase, experts from the crowdsourcing platform Prolific evaluated the design prototypes created without auxiliary systems (benchmark), the design prototypes later improved with the web-based representation, and the design prototypes improved with TESY. The experts rated their agreement with three statements *the design prototype perfectly fulfills the principle design task*, *the design prototype has high usability*, *the tasks can be solved with the design prototype* on 9-point Likert scales. Figure 8 shows the ratings in the form of boxplots. Additionally, in Table 2 the results of Mann-Whitney-U-Tests for selected designs are shown.

On the left third of Figure 8, the experts' ratings for *the design prototype perfectly fulfills the principle design task* are shown. Each boxplot represents the aggregated ratings on one class of design prototypes. The first class (light blue) are design prototypes by the baseline participants. Those design prototypes were not later improved. The second class are design prototypes created by the group that got the web-based representation of test data. Since this group improved their designs two boxplots are shown (dark yellow: initial design prototypes, light yellow: improved prototypes). In the third class, initial (dark red boxplot) and improved (lighter red boxplot) design prototype evaluations using TESY are shown. The two other thirds of Figure 8 follow the same schema, but since the respective group of participants did not create usability test tasks for the baseline designs, there is no baseline boxplot in the last third of Figure 8.

Table 2. Results of two-sample Mann–Whitney U tests for three different statements and our three treatments (*Base*, *Web*, *TESY*) rated by experts on Prolific. We hypothesized ex ante, that TESY would improve the ratings for *design task fulfillment*, *usability* and *task solvability*. We therefore conducted one-sided tests, and with "H_0: *location shift (Group A vs Group B) is less than or equals 0*". Effect size categories small (sm.), medium (md.), large (lg.) based on Cohen [14, p. 79 - 80], z-scores and r-values based on one-sided tests (see hypotheses).

Statements Rated by Experts on Prolific	**Group A**				**Group B**			**p**	**z**	**U**	**r**	**Significance, Cohen Effect Size**
	Web		TESY		Base	Web						
	T_0	T_1	T_0	T_1	T_0	T_0	T_1					
"The designs" perfectly fulfills the design task.		X			X			0.382	-0.301	3912.5	0.023	-
				X	X			0.096	-1.305	3587.0	0.103	-
				X			X	0.179	-0.921	3072.0	0.075	-
"The design has a high usability."		X			X			0.332	-0.435	3957.7	0.033	-
				X	X			0.071	-1.467	3637.0	0.115	-
				X			X	***0.017**	-2.120	2514.5	0.187	*5%, sm. to md.
"I could solve the tasks with the design."			X			X		0.074	-1.445	2339.5	0.127	-
				X			X	***0.026**	-1.941	3336.0	0.158	*5%, sm. to md.

Before we gathered the assessments from the experts on Prolific, we suspected that the quality of the design prototypes would be significantly better for selected comparisons. For example, we suspected that design prototypes that were improved in the second sessions would be rated higher by the experts - in comparison with the respective initial design prototypes, but also in comparison with the design prototypes of the unimproved baseline. Furthermore, we assumed that the direct embedding of usability test results with TESY would lead to better results. For the non-parametric Mann-Whitney U tests, we therefore decided beforehand to use one-sided tests.

When judging whether *how well the design prototypes fulfill the principle design task*, we find somewhat similar means and median values for the baseline designs and the initial designs for both treatments *web representation* and *TESY representation* (see Figure 8). For both mentioned treatments the improved design prototypess achieved higher means, and the treatment *TESY representation* resulted in a higher median. Though, the media values did not proof to be significantly higher in vice-versa comparisons for the design prototypes *fulfilling the principle design task* (Table 2).

Experts' ratings on the designs' usability have comparable means for the baseline, both initial and improved designs with treatment *web representation* and the initial designs with treatment *TESY representation*. However, we found statistically significant higher[1] expert ratings for the usability of the improved designs when comparing treatment *TESY representation* (Median: 7) with treatment *web representation* (Median: 6), $U = 2514.50$, $Z = -2.120$, $p < .05$, $r = 0.187$ (see Table 2).

[1] small to medium effect size, categories based on Cohen [14, p. 79-81]

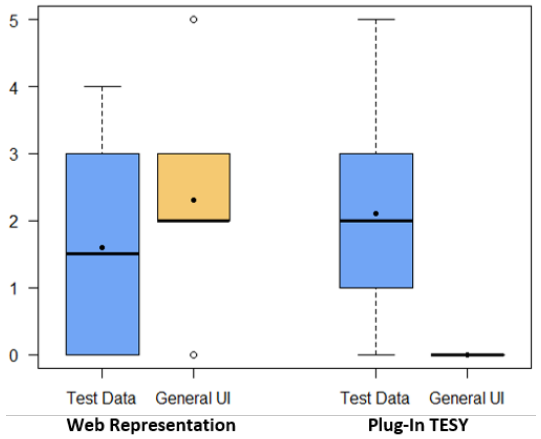

Fig. 9. Average number of self-expressed improvements per participant (y-axis) derived from usability test data (*Test Data*, light blue) and not data-based improvements (*General UI*, light orange) by each group in the third phase. Participants received test-data either *web-based* or via *TESY*. Black dots represent means.

When our experts rated whether they could *solve the tasks using the design prototypes* in both treatment groups *web representation* and *TESY representation* the initial designs show similar median and mean values (see third box in Figure 8). The treatment group *TESY representation* in contrast to group *web representation* achieved higher mean values when comparing the improved designs. Here, we found statistically significant higher[1] expert ratings for *task solvability* of the improved designs when comparing the treatment group *TESY representation* (Median: 8) with the treatment group *web representation* (Median: 8), $U = 3336.0$, $Z = -1.941$, $p < .05$, $r = 0.158$ (see Table 2).

5.2 Qualitative Results and Impact on Work Patterns

Participants received usability test data either web-based or integrated into the prototyping tool with TESY. As part of the survey, they were asked to report improvements made to their designs based on the usability test data. Based on the participants' self-reported improvements, one author of the paper analyzed the descriptions and categorized whether they described a change clearly explained as an improvement based on a tester's interaction or a general improvement of the user interfaces. Authors together discussed and resolved discrepancies. Participants using integrated test-data exclusively reported improvements where they provided clear links to usability test data sessions. Participants that received the test data web-based, reported slightly less improvements that they linked to usability test data sessions, but a significant amount of general, not data based, improvements of the user interfaces such as *"I have improved the visuals. Before I didn't like the colors. Now there is harmony, it looks good, and especially homogeneous"*. The focus area of both groups is displayed in Figure 9.

We assessed whether the improvements stated were conducted based on usability *test data* or independent improvements of the general user interface. The analysis of the self-expressed qualitative data also revealed that the majority of participants felt that they understood the usability test data, liked the analysis of the data, and two participants were even satisfied with it: *"Seeing users interact with their own ideas was very interesting. Seeing one's thoughts about the design confirmed with the test data was satisfying"* (Participant 3 in the improvement phase using TESY for data representation), another participant appreciated: *"trying to put myself in the tester's shoes and read*

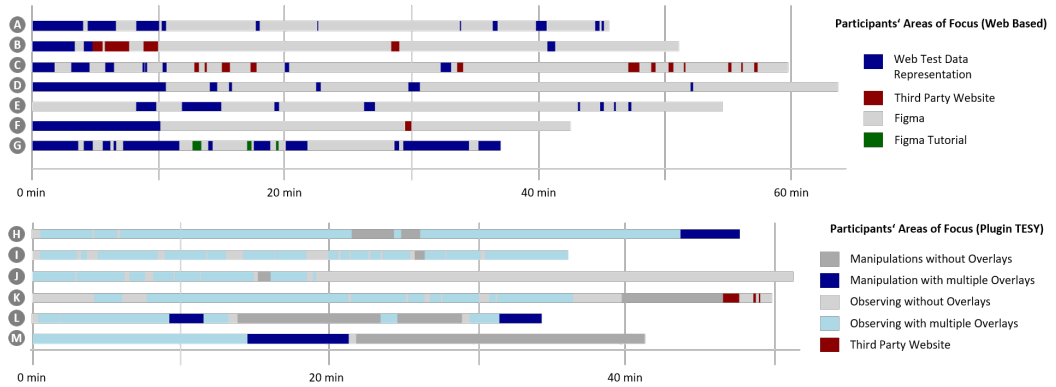

Fig. 10. **Upper half (A-G)**: participants' timeline using the web representation of usability test data. Here, blue areas represent times when only the website with test data was displayed, grey areas when just Figma was displayed, and dark green areas when participants re-watch the Figma tutorial.
Lower half (H-M): timeline of participants who used TESY to display results. Dark grey represents times when changes to the prototypes were made without displaying test data as an overlay, and dark blue times when changes were made with test data as an overlay. Light grey represents times when participants observed the prototypes without overlays, and light blue with overlays shown. Red areas describe times when a third-party website was displayed (e.g., to select icons for the GUIs). A-M symbolize the participants' sessions.

his views/troubles on my user interface from the lines." (Participant 4 in the improvement phase using TESY for data representation).

Furthermore, we evaluated the screen recordings of the participants to understand how participants used TESY and the web based test data representation. Due to a technical malfunction, not all screen recordings could be fully analyzed. Three screen recordings stopped during the experiments of participants who received the test data through the web representation, leaving seven screen recordings for analysis. Four screen recordings were interrupted for participants using TESY for data representation and could thus not be considered for this evaluation. For the following part, all numbers were calculated for the seven and six sessions that could still be analyzed. In order to analyze the screen recordings, one author went through and annotated each recording several times. Figure 10 shows the annotation results.

For the participants who viewed the usability test data on the website, it was annotated whether the website with test data, Figma, or another window was opened in full screen on the study computer screen (the participants had one monitor available in the study, no participant set a split screen). For participants using TESY for data representation, one author annotated whether the participants had chosen to display the test data as overlay data in Figma or hid the test data. Since for these participants, there was no switching between Figma and another window with test data; it was also annotated whether the participants were only observing the GUI prototypes (while moving their mouse) or actively making changes to the GUI prototypes (e.g., changing or adding text, images, or shapes, but also looking for assets within Figma, or working on the layers of the prototypes within Figma)[2].

[2]Since editing can involve minimal pauses during which nothing is actively changed, we have switched the labels from "making changes" to "observe," whenever no active changes to the prototype have been made in the previous 10 seconds. We switched labels from "observe" to "making changes" as soon as mouse or keyboard based changes were made.

On average, the participants who received the web-based test data spent 10m 09s with just the web representation open on the screen, 38m 51s with just Figma displayed, 11s re-watching the presented tutorial, and 43s on third-party websites (e.g., a website to get icons for the design tasks). Participants using TESY for data representation had the option to overlay test data from usability tests on their GUI prototypes, demonstrating their control and decision-making. On average, participants in this group spent 16m 58s using Figma without displaying test data as an overlay (9m 16s viewing the GUI prototypes, 07m 42s altering the GUI prototypes). For 26m 17s, participants chose to display the test data as an overlay (23m 36s observing the data, 02m 41s altering GUI prototypes while displaying test data as an overlay).

6 Discussion

In the following, we synthesize and discuss the findings of our study. On a quantitative side, while our findings show no significant influence on task load, creativity support, or system usability by using our plug-in TESY for integrating usability test data, the resulting designs designed with integrated test data were rated significantly high in terms of usability and task solvability. On a qualitative side, our evaluation shows that integrating usability test data with TESY leads to more time spent displaying test data and participant self-reporting, focusing on GUI improvements that can be linked to test data. We conclude the discussion by presenting design implications derived from the insights collected in our evaluation.

6.1 Interpreting Quantitative Results

Task Load, System Usability and Creativity Support Unfortunately, no significant differences could be found for NASA Raw TLX, SUS or CSI. We assume that more than the current sample size is needed to show significant effects. We decided to report the results so we could discuss them anecdotally. Looking at the absolute values, both NASA Raw TLX and SUS show the potential to reduce the perceived task load and increase the perceived usability for participants who used only Figma with and without plugins to create the usability tests (in the first phase) and for participants in the third phase where GUI prototypes were improved.

Expert Ratings of the GUI Prototypes Experts' evaluation of the GUI prototypes has shown that the GUI prototypes improved with TESY (i.e., the integrated presentation of the test data) were significantly better rated in terms of perceived usability and solvability of the usability tests. Here, the participants' focus on the test data described in the qualitative part could have solved the usability and test problems identified by the crowd-testers. As expected, generally improved prototypes (using TESY or the web representation) were rated better, on average, for fulfilling the design task, high usability, and task solvability. We found no significant difference in the ratings on fulfillment of the overall design task. However, we speculate that because the participants who received the test data disconnected made a higher percentage of general usability improvements, these changes also contributed to the fulfillment of the overall design task and, to some extent, to the usability but less to the actual solvability of the usability test tasks.

6.2 Interpreting Qualitative Results and Impact on Work Patterns

Participants' Areas of Focus. Our analysis of the displayed windows and test data overlays reveals interesting insights. Participants who received test data via the website looked at the test data primarily at the beginning of the sessions (participants A to G in Figure10). Except for participant G, all other participants with this treatment only looked at the test data very briefly later in the session (in the second half of the sessions, participants A-F never looked back at the test data for more than 2 minutes at times). Overall, it also shows that the participants spent more time in Figma

than looking at the test data (on average, 20.3% of the time with the test data website displayed and 77.9% with Figma without the test data website displayed). In contrast, in the group that received the test data in Figma, the participants chose, on average, 60.5% of the time to use Figma with test data displayed and 39.0% of the time to use Figma without test data displayed. However, it must be noted that whenever the participants actively made changes in the group that used TESY to present the results within Figma, the data was more likely to be hidden (the ratio of hidden test data to displayed test data while changes were being made is 2.86:1).

Participants who used TESY to display the usability test data within Figma spent considerably more time looking at the test data (average share in the sessions of 60.5% with TESY vs. 20.3% with the web representation). During this time, participants went through different usability test tasks and crowd-tester tasks. While three participants using TESY (H, L, M in Figure10) made direct changes, at the same time test data was displayed, some participants only changed when test data was not displayed, such as participants I and J in Figure 10. Interestingly, both participants made only minor changes directly related to feedback (i.e., changing the interactive links of components crowd-tests could not correctly use). The patterns found by analyzing the focus areas, suggest that participants made active use of the ability to display test data in Figma, spent more time viewing the data with this integration, and thus placed greater overall focus on the test data.

Data-based Improvements vs. Continuation of Design. Using TESY during design prototype improvements resulted in participants focusing on describing improvements that could be linked to the usability test data. In contrast, the participants who received the test data in the web-based interface focused more on describing non-test data-based improvements (such as creating additional subpages). We believe, this indicates that TESY enforced participants to follow a more data-driven approach in GUI prototype improvement. This self-expressed focus is consistent with the analysis of working patterns. The participants who received test data via the website spent a lot of time in Figma without constantly looking at the test data. We speculate that the lack of integration of the test data may have led to a higher percentage of participants making adjustments that they themselves did not see directly related to test data. The participants who received the data integrated in Figma had test data in their field of vision for a longer period and focused to a greater extent on improvements that they themselves described as related to errors in the usage patterns.

6.3 Design Implications

In Situ Integration of Usability Test Specification. Our experiment shows that novices can use an in situ integrated usability test specification tool, in the form of our plug-in TESY, to create usability test tasks directly in Figma while designing the GUI prototypes. Testers could then complete these tasks on a website set up expressly for testing the GUI prototypes. This website identically can be used by crowd-testers sourced on *Prolific* or *Amazon.com's Mechanical Turk*. The interface for creating tasks, with title, task, the criticality of the task, and description of the solution path (start frame in Figma, target frame in Figma, the path via other frames), has proven feasible so that even beginners can create usability tests with Figma.

In Situ Integration of Usability Test Data. The results' analysis shows that integrating usability test results in situ into GUI prototyping tools, such as Figma, in contrast to having this data presented on disconnected websites has significant advantages in two areas when used by novice designers. Experts rated design prototypes improved with TESY (i.e., with the integrated test data) significantly higher for general usability and solvability of the usability test task than design prototypes improved with the web-based representation of usability test tasks. Furthermore, our analysis of working patterns underscores a stark contrast. Participants who received usability test data integrated with TESY focused their efforts on making improvements that can be completely

linked to the test data. In contrast, those who received usability test data via the web representation tended to focus more on improvements that to a higher degree are described as improving the the general, not directly data related, usability. We recommend an in situ integration of usability test specification and usability test data into dedicated prototyping tools to researchers and practitioners, to support novice users' focus on the test data.

Careful Use of the Screen Space. We have implemented TESY with the goal of taking up as little space as possible so that plenty of area is available for drafting the design prototypes. Nevertheless, one of the most frequent improvement suggestions was the wish to be able to minimize the plug-in: *"The plug-in cannot be minimized. This means that the window is always open and therefore in the way."* (Participant 8 in the improvement phase using TESY for data representation). Since we were already aware of the users' wish to use the available space effectively, we made two design decisions to support the effective use of screen space. We had described in detail that the plug-in can be moved to the bottom of the screen. In addition, we took care in the design to store the data in the local storage of Figma so that after closing and reopening the plug-in, the entered data is available again. Since participants still commented towards better use of the screen space, future work in this area should explicitly implement a minimization function. We recommend to include a minimized state, for novice users, with a reduced and compact design, that can still hint at the core functionalities so that users can quickly jump back to the annotation, the task specification, and the display of the test data.

From Manual to Automated Interpretations of Test Data. Although most participants could successfully specify usability tests with TESY and make improvements based on the test data in the following steps, some participants reflected on their abilities and expressed limitations. One participant noted, *"As a layperson in graphic design, it was generally difficult to draw improvement ideas from the mouse movement data, even though it was presented."* (Participant 10 in the improvement phase using TESY for data representation). We believe that many unskilled designers may have problems interpreting the data. The annotation feature of TESY and the test data already lays the foundation to interpret the test data in an automated way. Future designs of automated assistants for in situ usability test integration may place a particular emphasis on automated interpretation of the data. In the context of unskilled designers, automated interpretation could be combined with explanations so that a better understanding of necessary improvements emerges. We found another opportunity for automated analysis derived from a suggestion from a participant: *"After moving the entire frames, the test results are no longer positioned correctly. This makes it difficult to correctly understand the movements and clicks."* (Participant 8 in the improvement phase using TESY for data representation). At a first glance, integration into the prototyping tool is a disadvantage as soon as the design prototype is changed because the interaction data no longer fits the current design. The same disconnect happens if the interaction data is presented with the web-based interface. However, automation could present an opportunity here. If some interactions can be clearly assigned to a component (e.g., near-miss clicks on a button), a real-time recalculation could be carried out when the component is moved, and the incorrect clicks could be moved in the design in real-time. We suggest that future research address the topic of dynamic click recalculation to identify opportunities and weaknesses of the approach.

7 Limitations And Future Work

In the following we would like to summarize limitations of our study and link them to future research directions.

Sample Size and Experimental Lab Environment. One limitation of our study, represents the sample size of the first, second and third phase of the study and its execution in an experimental lab

environment. Additionally, for the analysis of screen recordings, not all participant session could be considered due to a technical malfunction. Studies with comparable study designs have used similar sample sizes (e.g., [33, p. 200], where Cohen's $d = 1.09$ was found). Furthermore, all participants (with an exception to the experts on Prolific) were recruited from an university panel. Sampling the participants from a university panel may have introduced a tendency for specific characteristics (e.g., in terms of age, educational level, and income) and thus may limit the generalizability of our results. Thus, future research should scale sample size and in combination perform a real-world field study with TESY. Since TESY is implemented as a Figma plugin, a distribution to the large Figma user community is also possible.

Novice Designer vs. Trained Designers. With TESY, we have evaluated a system that enables significantly better implementation of improvements to design prototypes, especially for novice designers. However, this also resulted in some limitations. A system for experts may, of course, have different requirements. In addition, we found that participants in our study, while most had similar little prior knowledge of creating design prototypes, could handle the prototyping tool Figma with different skill levels. The participants' struggle with Fimga is not only reflected in the mediocre SUS and the NASA Raw TLX ratings for our control group (only Figma was rated here), but also in statements from individual participants, such as: *"The toolbar selection [of Figma] is too complex for someone who has not been using such programs before. What do I select?*" (Participant 4 in the improvement phase using the web-based representation). When interpreting the test data, it was also shown that some participants were better able to deal with usability test data than others. One participant who used TESY stated no general problems understanding the usability test data: *"[TESY] was quite easy to use for a layman and it was very easy to trace back where there were problems in the usability of my created program via the displayed mouse movements and clicks."* (Participant 4 in the improvement phase using the TESY). Another participant generally had difficulty understanding the usability test data: *"Multiple clicks were made on empty areas of the website. I could not interpret this."* (Participant 5 in the improvement phase using the web-based representation). Future research could, therefore, examine the extent to which the supposedly homogeneous group of unskilled designers needs personalization in the presentation and assistance in interpreting usability test data. We also recommend that future work may identify additional differences in requirements between beginner and expert designers and provide TESY with a beginner mode and an expert mode.

Realistic Nature of the Design Task. When creating the study, we explicitly decided that the participants were not allowed to use any images, besides icons from a predefined library with a fitting license, from the Internet in their design prototypes. On the one hand, we wanted to prevent participants from spending too much time looking for the right images, and on the other hand, copyrights would have had to be clarified for each image inserted. We instructed participants to use placeholders instead to still make these areas interactice as as click areas. However, the missing images did not contribute positively to the usability assessment by the experts on Prolific, as they significantly deviated from the design prototypes typically created in practice. In order to be even closer to practice with the tasks in subsequent studies, the question arises as to whether the participants can be provided with a usable and cataloged collection of images before the study to make the task even more realistic. Additionally, while using a task in a domain familiar to the participants, using just one task might limit our findings in regard to different contexts and domains. As mentioned, TESY ideally would be evaluated in a realistic field experiment with multiple different real-world design tasks and prototypes.

8 Conclusion

The creation of GUI design prototypes is becoming more and more important. In parallel, asynchronous usability testing has been recognized as a very efficient method to test and on this basis to improve design prototypes. However, the creation of usability test specifications and data evaluation is decoupled from the design prototypes in their prototyping environment. This paper presents TESY, a usability test-driven prototyping assistant for novice designers. TESY first of all allows defining usability tests with names, tasks, the relevance of the test, start frame, end frame, and interactive components to be created directly in prototyping tools such as Figma. In addition, TESY allows components to be annotated in the prototyping tool, which, among other things, leads to a better understanding of the different components in general and in particular to usability testing. The information captured by TESY can be used directly to create and carry out asynchronous usability tests. Subsequently, the test results can be imported into TESY. TESY allows users to visualize the collected test data directly in the prototyping tool. In our study, design prototypes improved with TESY were rated significantly higher by experts for general usability and solvability of usability test task when compared to design prototypes improved with the web-based representation of usability test tasks. According to the participants, the use of TESY's function of integrating test data into the prototyping tool, led to data-driven and more task-focused design improvements. Participants who received the test data decoupled vastly improved the general usability or expanded the design with new content. Our evaluation demonstrates that TESY has positive impact on its users as well as the resulting outcomes. The future of TESY holds promising avenues for further refinement, especially with regards to more advanced analysis of the collected data as well as automated interpretation.

References

[1] 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry* (0 ed.), Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). CRC Press, 207–212. doi:10.1201/9781498710411-35

[2] Carolina Abrantes, Óscar Mealha, Diogo Gomes, João Paulo Barraca, and Carlos Viana-Ferreira. 2022. Analyzing and Visualizing the Criticality of Issues from Usability Tests. *J. Usability Studies* 17, 2 (June 2022), 65–84.

[3] Adobe Inc. 2023. Adobe Illustrator. https://www.adobe.com/de/products/illustrator.html

[4] Salvatore Andolina, Hendrik Schneider, Joel Chan, Khalil Klouche, Giulio Jacucci, and Steven Dow. 2017. Crowdboard: Augmenting In-Person Idea Generation with Real-Time Crowds. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. Association for Computing Machinery, New York, NY, USA, 106–118. doi:10.1145/3059454.3059477

[5] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. 2007. What Happened to Remote Usability Testing? An Empirical Study of Three Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1405–1414. doi:10.1145/1240624.1240838

[6] Balsamiq SRL. 2023. Balsamiq Wireframes. https://balsamiq.com/wireframes/

[7] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (July 2015), 85–94. doi:10.1145/2791285 New York, NY, USA, Association for Computing Machinery.

[8] Lara Bertram and Markus Dahm. 2022. Conceptual design and implementation of an automated metrics and model-based usability evaluation of UI prototypes in Figma. In *Mensch und Computer 2022 - Workshopband*, Karola Marky, Uwe Grünefeld, and Thomas Kosch (Eds.). Gesellschaft für Informatik e.V., Bonn, 1–6. doi:10.18420/muc2022-mci-ws06-264

[9] Anders Bruun, Peter Gull, Lene Hofmeister, and Jan Stage. 2009. Let Your Users Do the Testing: A Comparison of Three Remote Asynchronous Usability Testing Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1619–1628. doi:10.1145/1518701.1518948

[10] Anders Bruun and Jan Stage. 2012. The effect of task assignments and instruction types on remote asynchronous usability testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 2117–2126. doi:10.1145/2207676.2208364

[11] Erin A. Carroll, Celine Latulipe, Richard Fung, and Michael Terry. 2009. Creativity factor evaluation: towards a standardized survey metric for creativity support. In *Proceedings of the seventh ACM conference on Creativity and cognition*. ACM, Berkeley California USA, 127–136. doi:10.1145/1640233.1640255

[12] José C. Castillo, H. Rex Hartson, and Deborah Hix. 1998. Remote Usability Evaluation: Can Users Report Their Own Critical Incidents?. In *CHI 98 Conference Summary on Human Factors in Computing Systems (CHI '98)*. Association for Computing Machinery, New York, NY, USA, 253–254. doi:10.1145/286498.286736

[13] Yan Chen, Maulishree Pandey, Jean Y. Song, Walter S. Lasecki, and Steve Oney. 2020. Improving Crowd-Supported GUI Testing with Structural Guidance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376835

[14] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge, New York. doi:10.4324/9780203771587

[15] J. F.C. de Winter and D. Dodou. [n. d.]. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012). ([n. d.]). doi:10.7275/BJ1P-TS64 Publisher: University of Massachusetts Amherst.

[16] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A Mobile App Dataset for Building Data-Driven Design Applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. Association for Computing Machinery, New York, NY, USA, 845–854. doi:10.1145/3126594.3126651

[17] Biplab Deka, Zifeng Huang, Chad Franzen, Jeffrey Nichols, Yang Li, and Ranjitha Kumar. 2017. ZIPT: Zero-Integration Performance Testing of Mobile App Designs. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. Association for Computing Machinery, New York, NY, USA, 727–736. doi:10.1145/3126594.3126647

[18] Figma, Inc. 2023. Figma. https://www.figma.com

[19] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3290605.3300619

[20] Google LLC. 2023. Material Design. https://m3.material.io/

[21] John D. Gould and Clayton Lewis. 1985. Designing for Usability: Key Principles and What Designers Think. *Commun. ACM* 28, 3 (March 1985), 300–311. doi:10.1145/3166.3170

[22] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. doi:10.1177/154193120605000909

[23] H. Rex Hartson, José C. Castillo, John Kelso, and Wayne C. Neale. 1996. Remote Evaluation: The Network as an Extension of the Usability Laboratory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*. Association for Computing Machinery, New York, NY, USA, 228–235. doi:10.1145/238386.238511

[24] Saskia Haug, Ivo Benke, Daniel Fischer, and Alexander Maedche. 2023. CrowdSurfer: Seamlessly Integrating Crowd-Feedback Tasks into Everyday Internet Surfing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. doi:10.1145/3544548.3580994

[25] Saskia Haug and Alexander Maedche. 2021. Feeasy: An Interactive Crowd-Feedback System. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 41–43. doi:10.1145/3474349.3480224

[26] Lena Hegemann, Niraj Ramesh Dayama, Abhishek Iyer, Erfan Farhadi, Ekaterina Marchenko, and Antti Oulasvirta. 2023. CoColor: Interactive Exploration of Color Designs. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 106–127. doi:10.1145/3581641.3584089

[27] Jason Hong. 2011. Matters of design. *Commun. ACM* 54, 2 (Feb. 2011), 10–11. doi:10.1145/1897816.1897820

[28] Kasper Hornbæk and Erik Frøkjær. 2008. Making use of business goals in usability evaluation: an experiment with novice evaluators. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Florence Italy, 903–912. doi:10.1145/1357054.1357197

[29] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-Situ Story Ideation with the Crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376715

[30] Bernard J. Jansen. 1998. The graphical user interface. *ACM SIGCHI Bulletin* 30, 2 (April 1998), 22–26. doi:10.1145/279044.279051

[31] Felix Kretzer and Alexander Maedche. 2023. Making Usability Test Data Actionable! A Quantitative Test-Driven Prototyping Approach. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–6. doi:10.1145/3544549.3585659

[32] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. 2015. Apparition: Crowdsourced User Interfaces that Come to Life as You Sketch Them. In *Proceedings of the 33rd Annual ACM Conference*

on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 1925–1934. doi:10.1145/2702123.2702565

[33] Chunggi Lee, Sanghoon Kim, Dongyun Han, Hongjun Yang, Young-Woo Park, Bum Chul Kwon, and Sungahn Ko. 2020. GUIComp: A GUI Design Assistant with Real-Time, Multi-Faceted Feedback. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376327

[34] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D. O'Keefe, and Walter S. Lasecki. 2018. Exploring Real-Time Collaboration in Crowd-Powered Systems Through a UI Design Tool. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018). doi:10.1145/3274373

[35] Sang Won Lee, Yujin Zhang, Isabelle Wong, Yiwei Yang, Stephanie D. O'Keefe, and Walter S. Lasecki. 2017. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping of Interactive Interfaces. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. Association for Computing Machinery, New York, NY, USA, 817–828. doi:10.1145/3126594.3126595

[36] Di Liu, Randolph G. Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10. doi:10.1002/meet.14504901100

[37] Marvel Prototyping Ltd. 2023. Ballpark. https://ballparkhq.com/

[38] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 473–485. doi:10.1145/2675133.2675283

[39] MAZE.DESIGN LIMITED. 2023. Maze. https://maze.co

[40] James D. McKeen, Tor Guimaraes, and James C. Wetherbe. 1994. The Relationship between User Participation and User Satisfaction: An Investigation of Four Contingency Factors. *MIS Quarterly* 18, 4 (1994), 427–451. http://www.jstor.org/stable/249523

[41] Michael Nebeling, Maximilian Speicher, and Moira C. Norrie. 2013. CrowdStudy: general toolkit for crowdsourced evaluation of web interfaces. In *Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '13)*. Association for Computing Machinery, New York, NY, USA, 255–264. doi:10.1145/2494603.2480303

[42] Jakob Nielsen. 1993. Usability Engineering. In *Usability Engineering*, Jakob Nielsen (Ed.). Morgan Kaufmann, San Diego, 1–21. https://www.sciencedirect.com/science/article/pii/B9780080520292500048

[43] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. DesignScape: Design with Interactive Layout Suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1221–1224. doi:10.1145/2702123.2702149

[44] Jonas Oppenlaender, Thanassis Tiropanis, and Simo Hosio. 2020. CrowdUI: Supporting Web Design with the Crowd. *Proc. ACM Hum.-Comput. Interact.* 4, EICS (June 2020). doi:10.1145/3394978

[45] Srishti Palani, David Ledo, George Fitzmaurice, and Fraser Anderson. 2022. "I Don't Want to Feel like I'm Working in a 1960s Factory": The Practitioner Perspective on Creativity Support Tool Adoption. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3491102.3501933

[46] Prolific Academic Ltd. [n. d.]. Prolific. www.prolific.co

[47] Christoph Schneider and Terence Cheung. 2013. The Power of the Crowd: Performing Usability Testing Using an On-Demand Workforce. In *Information Systems Development*, Rob Pooley, Jennifer Coady, Christoph Schneider, Henry Linger, Chris Barry, and Michael Lang (Eds.). Springer New York, New York, NY, 551–560.

[48] Studio XID, Inc. 2023. ProtoPie. https://www.protopie.io/

[49] Useberry User Testing Technologies P.C. 2023. Useberry. https://www.useberry.com/

[50] USER TESTING INC. 2023. UserTesting. https://www.usertesting.com/

[51] Karel Vredenburg, Ji-Ye Mao, Paul W. Smith, and Tom Carey. 2002. A Survey of User-Centered Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. Association for Computing Machinery, New York, NY, USA, 471–478. doi:10.1145/503376.503460

[52] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1433–1444. doi:10.1145/2531602.2531604

Received January 2024; revised July 2024; accepted October 2024