Towards Better User Studies in Computer Graphics and Vision

Zoya Bylinskii
Adobe Research
bylinski@adobe.com

Laura Herman
Adobe Inc.
lherman@adobe.com

Aaron Hertzmann
Adobe Research
hertzman@adobe.com

Stefanie Hutka
Adobe Inc.
stefanie.hutka@gmail.com

Yile Zhang
University of Washington
yz278@uw.edu

Contents

Towards Better User Studies in Computer Graphics and Vision

Zoya Bylinskii[1], Laura Herman[2], Aaron Hertzmann[1], Stefanie Hutka[2] and Yile Zhang[3]

[1] *Adobe Research, USA; bylinski@adobe.com, hertzman@adobe.com*
[2] *Adobe Inc., USA; lherman@adobe.com, stefanie.hutka@gmail.com*
[3] *University of Washington, USA; yz278@uw.edu**

ABSTRACT

Online crowdsourcing platforms have made it increasingly easy to perform evaluations of algorithm outputs with survey questions like "which image is better, A or B?", leading to their proliferation in vision and graphics research papers. Results of these studies are often used as quantitative evidence in support of a paper's contributions. On the one hand we argue that, when conducted hastily as an afterthought, such studies lead to an increase of uninformative, and, potentially, misleading conclusions. On the other hand, in these same communities, user research is underutilized in driving project direction and forecasting user needs and reception. We call for increased attention to both the design and reporting of user studies in computer vision and graphics papers towards (1) improved replicability and (2) improved project direction. Together with this call, we offer an overview of methodologies from user experience research (UXR), human-computer

---

*While on internship at Adobe.

---

interaction (HCI), and applied perception to increase exposure to the available methodologies and best practices. We discuss foundational user research methods (e.g., needfinding) that are presently underutilized in computer vision and graphics research, but can provide valuable project direction. We provide further pointers to the literature for readers interested in exploring other UXR methodologies. Finally, we describe broader open issues and recommendations for the research community.

1

Introduction

Most research in computer graphics and image synthesis produces outputs for human consumption. In many cases, these algorithms operate largely automatically; in other cases, interactive tools allow professionals or everyday users to author or edit images, video, textures, geometry, or animation. For example, photo manipulation algorithms allow artists and casual photographers to modify images for expression and visual communication; geometry synthesis algorithms allow artists to create geometry for video games and movies, to facilitate architectural and industrial design; material models can then be used to texture the geometries; image restoration algorithms, such as super-resolution and colorization, aim to produce visually plausible and appealing images. Likewise, many synthesis algorithms published in computer vision are also designed for human consumption, including generative AI, image enhancement, image stylization, neural rendering, and 3D capture of faces and bodies. **When the tools or outputs are meant for user consumption, at what point in the project should users be brought in to evaluate them, and how can the results of user studies further benefit the research?**

We have recently seen a proliferation of research papers in computer vision and graphics venues reporting "user studies" in which crowdworkers rate algorithm outputs, executed at the end of the project timeline, as an afterthought or in response to the review process. On the one hand, when conducted hastily and without sufficient attention to the study design choices, replicability of the published study results can suffer. We encourage authors and paper reviewers alike to evaluate whether and when a user study is necessary, and to avoid asking for—or running—perfunctory studies that do not affect the paper's conclusions or project directions. On the other hand, the true benefit of user studies lies in having them shape the evolution and strategy of a project, or as is the case with foundational research, even the initial project direction. When conducted at the very end, researchers leave no space or time for the results of the user studies to lead to meaningful project improvements or iterations.

These considerations are particularly timely with regards to the recent explosion of generative AI technologies. The gap from research iterations to consumer-facing products has shrunk, and users are increasingly being put in front of powerful image and text generation technologies with enormous ethical, legal, and societal implications. In these cases, the types of computational benchmarks common to other facets of vision and graphics research are less relevant, and instead, the focus turns to user behavior, reactions, and interactions with the technology. Here the opportunities for user research are to assess user needs and to forecast user behavior and reception early on and regularly during the model development lifecycle.

Assuming that researchers want their algorithms to be used in the real world, developing useful tools often requires talking to real users. However, getting meaningful feedback is very difficult and may require specialized expertise. This discipline of understanding the user, their needs, and feedback is called *user research*, and was born out of the intersection of psychology and human-computer interaction, pioneered by electrical engineer and psychologist Don Norman. Many technology companies employ user experience researchers, or UXRs (including coauthors on this work). While we urge researchers to collaborate with experts—such as UXRs, HCI researchers, or human perception

scientists—this is not always possible. Further, some models and applications (e.g., generative AI) may require testing on a larger user base than would be tractable for qualitative methods. For these reasons, **this monograph offers a guide and introduction to user research methodologies relevant for graphics and vision researchers**. For further reading, we provide pointers to key resources on user research, and the terminology for talking about user research that can help navigate those resources. This monograph draws on our own academic and industrial experience with user research, within computer graphics, vision and other areas. In providing this background, we hope to expand vision and graphics researchers' repertoire of user study methodologies for gaining different types of insights throughout the project lifecycle.

We categorize user research methods into three buckets: *Output Evaluation* (Section 3.1), used to evaluate the outputs of an algorithm or compare outputs between algorithms; *Interface Evaluation* (Section 3.2), used to evaluate how an interactive tool can support or augment a user's typical workflow or otherwise facilitate task completion; and lastly, *Foundational Research* (Section 3.3), performed before any tool or algorithm has been built, to help guide design and development to meet real user needs. This last type of user research is rare in vision and graphics research, but more common in HCI and corporate product development. We describe techniques for designing effective evaluations, getting more information from studies, and avoiding common pitfalls that may invalidate results or hinder replicability.

In this monograph, our goal is to elevate the role of user studies in graphics and vision research. We argue that they should be treated with the same care and rigor expected of other parts of the research, and in doing so, can directly shape the project direction. We close with a maxim to keep in mind: **bad user research leads to bad outcomes**, and we discuss ways that flawed user studies can mislead or misguide research and product development. We hope this monograph will help researchers perform better user studies, leading to useful evaluations and new insights that can inform and inspire their research.

2

The Start of User Studies in Graphics and Vision

Computer graphics research has long developed algorithms and interfaces for creating and manipulating images, videos, textures, geometry, and animation. Computer vision venues now also actively publish papers on these topics. While the graphics community has long discussed the need for rigorous evaluation for more aesthetic and output-focused work [1], [45], evaluations of aesthetic and interactive techniques appeared only rarely in the first four decades of SIGGRAPH's history. For example, in 2008, just before the introduction of crowdsourcing, only 3 of the 103 papers published at SIGGRAPH 2008 performed studies with human evaluation, and none of these papers reported user evaluations of interfaces, or any user feedback. Historically, only a few types of graphics papers leveraged user research, including perceptual studies for rendering [23], [61], [96], [106] and animation [62], [72], [73], as well as, occasionally, for evaluating interactive systems [99], [104].

Beginning in 2008, **crowdsourcing drove the graphics and vision fields toward widespread use of "user studies".** Papers in the HCI literature showed that Amazon Mechanical Turk (MTurk) could be used for user evaluations [52] and perceptual studies [43]. This work inspired new research using MTurk to evaluate algorithm outputs,

notably O'Donovan *et al.* [71]. Likewise, computer vision papers in 2008 showed that human computation could be used for data labeling [91], [101], [102], [107], a purpose for which it became widespread [29], [54]; it is also used for evaluating predicted labels [7], [19], [56], [75]. Crowdwork became widespread in other fields; for example, a highly-cited psychology paper by Buhrmester *et al.* [13] pointed out that MTurk could supplant the usual practice in psychology of using one's own undergraduate students as the data source, a practice that may have produced many misleading and biased results [49]. In addition to crowdsourcing platforms, it is not uncommon to see the use of convenience sampling or friend-sourcing (involving friends or labmates) in the evaluations of vision and graphics algorithm outputs.

Crowdsourcing made it easy to quickly and efficiently gather large numbers of human-provided labels and evaluations that would have otherwise been very difficult to acquire. While providing a way to collect anonymous user feedback at scale, it has frequently been done without the careful attention to study design and methodology that is practiced in the user experience and perception communities (e.g., properly balanced conditions, statistical measurements of confidence and effect size, careful consideration of confounding factors and sources of bias, etc.). This lack of rigor contrasts with the careful attention paid to benchmark datasets and other types of quantitative evaluation which are often prioritized in vision and graphics papers. Performing effective studies online requires careful thought and effort [26]. The true potential of user studies to shape tool development and project direction is underutilized. Moreover, the field has yet to make use of the broader range of user research techniques available, which could offer insights throughout the project lifecycle.

In the next sections, we categorize user study designs into those most applicable to graphics and vision researchers for evaluating outputs, evaluating interfaces, and ideating project directions in the first place. Where applicable, we include examples of papers from the literature that have used the different types of study designs discussed.

3

User Research Methods for Graphics and Vision

In computer graphics and vision research, user studies are commonly used for evaluating one of two contributions: the final output of a model or algorithm, which we refer to as *output evaluation*, or the user interface itself, which we refer to as *interface evaluation*. For instance, the typical survey deployed on Amazon's Mechanical Turk (MTurk) or a similar crowdsourcing platform that asks participants to rate which of a set of outputs (images, videos, textures, etc.) are "better," is an example of output evaluation (see Section 3.1.1 for recommendations for running these types of studies).

One way to think about user study methodologies is to consider where they occur on the research & development timeline (Figure 3.1). This is common practice in corporate product development, but applies well to research. For instance, surveys represent just one possible study design, and together with other output and interface evaluations, occur near the end of the research & development timeline. *Foundational research* (including methods like needfinding interviews), on the other hand, occurs before any tool development takes place and is used to define the problem statement and guide development. While rarer in graphics and vision papers, we advocate its wider use, as it can help
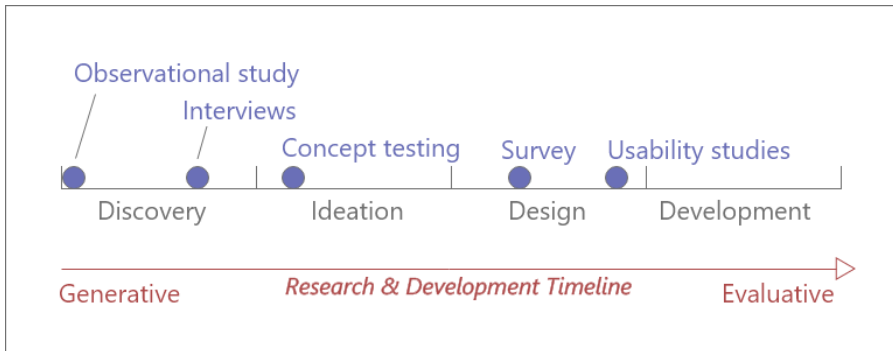
Figure 3.1: Overview of common methodologies from user experience research relevant to vision and graphics researchers, spanning the research & development timeline. At the beginning of this timeline are generative methods like observational studies and interviews that help to define the problem and guide the research. At the other end of this timeline are evaluative methods like surveys and usability studies for assessing the final results.

lead toward more impactful projects. The rest of this section is devoted to describing and providing recommendations for output evaluation (3.1), interface evaluation (3.2), and foundational research (3.3), but we first introduce some terminology.

User experience research (UXR) categorizes methodologies using the following axes: qualitative/quantitative, attitudinal/behavioral, and generative/evaluative. *Qualitative* research involves collecting non-numerical data to understand behaviors or opinions ("how or why do people do something?"). Conversely, *quantitative* research involves collecting and analysing numerical data ("what percentage of people in a population do something?"). *Attitudinal* approaches focus on what users say, whereas *behavioral* approaches focus on what users do. Generative approaches collect information to help form the foundation of the research and development process. Questions such as, "whom are we building this tool for?", and "what user problems are we solving?" are commonly addressed with *generative* approaches. In contrast, *evaluative* approaches collect information to refine an existing concept or tool. Discussing study methodologies in terms of these axes makes explicit what a given user study is measuring (e.g., user attitudes or user behaviors?) and for what purpose (e.g., to guide the development of a tool or to evaluate its usability once built?).

What do you hope to learn? In deciding what types of user studies to run for a research project, the critical questions are: what are your goals, what do you hope to learn, and what data do you hope to gather? If the goal is to learn from existing user workflows, then a *generative* approach, such as an *observational study* or *interview* is appropriate. If the goal is to report which algorithm is better, a *quantitative survey*, performed in the *evaluative* phase of the research & development cycle, may suffice. But for additional insights about *why* users made the selections they did, trade-offs between algorithms, or failure modes of an algorithm, *qualitative* approaches like *interviews* or *focus groups* should be considered. User research can be carried out throughout the duration of research & development, to test evolving assumptions about user workflows and continue to iterate on tool design. Doing so can also help reduce surprises during final evaluation.

Should you perform a study at all? Before embarking on user studies for a paper, we recommend thinking through the study and whether it will provide genuinely useful information. For example, if the study aims to evaluate image outputs, will the study tell you something you cannot tell from looking at the images yourself? If the study will not be performed carefully and rigorously, will the results be meaningful and replicable? If crowdworkers will be used, will their responses be representative of the target users of the tool/algorithm? Does the technique even merit evaluation, e.g., for early-stage experimental designs, where a meaningful baseline comparison might not possible [38]? Likewise, paper reviewers should ask themselves similar questions before making demands on the authors. We discuss these considerations in more depth in Section 6, where we argue that it is better to perform no user study at all, than to perform a poorly-designed or perfunctory study.

3.1 Output Evaluation

We use *output evaluation* to refer to evaluations of the final products of algorithms such as synthesized or modified images, textures, animations, or videos, e.g., algorithms designed for inpainting, tone adjustment, image stylization, or texture synthesis. Examples of output evaluations of generative AI models include [30], [113]. Both leverage psychophysics-

inspired experiments that display a sequence of images to participants and have them evaluate whether each image is real/fake, while varying image presentation times to compare output quality across models (i.e., an image detected as fake in a shorter presentation time represents a poorer quality model result). Note that this study design does not however evaluate *why* (what aspect of) an output is perceived as fake.

Output evaluation can be carried out as an interview, a questionnaire, a survey, or an observational study, although surveys are most common in computer vision and graphics papers. In UXR terminology, a *survey* involves asking study participants to rate or compare alternatives, in order to generate metrics about user ratings. It is a *quantitative*, *attitudinal*, and *evaluative* study design.

Properly designed surveys improve the likelihood that the study is replicable. Beyond the high-level recommendations that we provide below, good practices for survey design can be found in Rea *et al.* [82] and Kuniavsky [55]. However, there are also whole courses on survey design and psychometrics that UX researchers often take as part of formal training or graduate degree programs, so we recommend consulting with an expert if possible.

3.1.1 Study Recommendations

We now describe several interface designs for rating images, several of which are summarized in Figure 3.2. Consider the case of two image generation algorithms, algorithm A and algorithm B, that need to be compared. This requires a selection of images on which to compare these algorithms, and an interface that will present images to participants, asking them to evaluate the results, for instance by clicking on the image of the pair they prefer (Figure 3.2d). Participants will be asked to make some number of judgements in sequence. This type of evaluation study is most typically run by graphics and vision researchers on a crowdsourcing platform, and the results are tallied to report what proportion of participants preferred algorithm A's outputs. This common design is also referred to as: two-alternative forced choice (2AFC), A/B testing [53], and randomized controlled trials in science, so similar study guidelines apply.
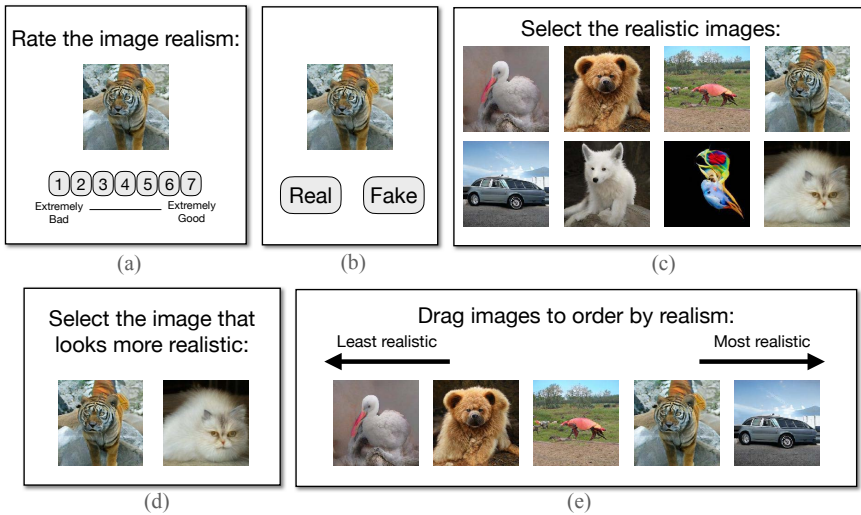
Figure 3.2: Sample user interface designs for output evaluations, ranging from single-image assessments and ratings, to A/B comparisons, and multi-image evaluations.

When a larger number of outputs or algorithms need to be compared, a pair can be sampled on each trial (e.g., [59], [85], [109]), or participants may be presented with three or more options simultaneously and asked to order or rank them (Figure 3.2e). The average rank assigned to images from a particular algorithm can be used as a summary score (e.g., [35]). Another option is showing multiple outputs at once, from any number of algorithms, asking participants to select all the images that match a particular criterion of quality (Figure 3.2c, e.g., [110]). During analysis, the proportion of images chosen from each algorithm can be used to compare the algorithms (e.g., [57]). Alternative designs include showing each of the options one by one or all at once, asking participants to rate each option on a Likert scale (Figure 3.2a, e.g. Rosala [87]), or to make a simple judgement for each image (e.g., real/fake like in Figure 3.2b as in [30], [113]). Ratings can be averaged across participants and used to rank the algorithms (e.g., [14]). Below we describe techniques for improving the quality of these studies, increasing the likelihood they give useful, replicable results, and avoiding common pitfalls.

Select a fair sample of outputs; the more, the better. To fairly compare two algorithms, example outputs should be randomly selected, rather than hand-picked. The exception is if different edge cases need to be explicitly represented (e.g., an equal distribution of natural scenes, cluttered scenes, portraits, etc.). The image selection criteria should be clearly reported for replicability. A *within-subjects* study design would be one in which each participant sees all conditions, and requires a different analysis than a *between-subjects* design that shows different conditions to different participants (e.g., [58], [59], [113]). Across evaluators, it is important to shuffle the sequence of presented images to avoid any ordering effects (e.g., learning/ramp-up, context effects, fatigue, etc., Elmes *et al.* [32]).

Anonymize the study. Whenever possible, study participants should not know which of the outputs is the authors' method. Otherwise, participants may respond with the answers they think the researchers want, consciously or not, which is known as the "good subject effect" (e.g., Nichols *et al.* [65]). Do not label outputs with names like "our method" or "existing method". Participants can be biased by power dynamics (i.e., the researcher holds power by running the study), researchers using language to prime participants (e.g., "how much do you like this tool I built yesterday?"), and researchers and participants' relationship (e.g., if both work in the same lab or company).

Consider task presentation. It is important to randomize the presentation order of the algorithms: if, say, in a side-by-side comparison the baseline is always shown on the left as image A and the new algorithm as image B, participants may infer that image B is the new method and prefer it; or they may even get in the habit of just clicking B rather than comparing each image pair separately. Other presentation aspects such as the size of the images on the screen, their distance to each other, etc. may influence participant responses. Piloting the study with a few different settings may help spot these potential confounds early, and provide an estimate of participants' answer variability that is due to study design, independent of algorithm performance.

Add checks throughout. Remote crowdworkers may click randomly to speed through a trial and get paid, rather than answering questions in good faith. Moreover, recent evidence indicates a substan-

tial decrease in MTurk data quality in the past few years, likely due to an increase of bad actors and bots [18]. For these reasons, it is important to build in checks throughout the study design (e.g., Sauro [92]). Checks (also called *sentinels* or *validation trials*) are trials for which there is an objectively correct answer that can be used to filter out poor-quality data (example validation conditions described in Lu *et al.* [59], [58], and Wang *et al.* [109]). For instance, in the case of pairs of generated images, one image of the pair can be an intentionally and objectively poor quality result. During analysis, data from participants that failed some preset number of the checks can be discarded, assumed to be generated by participants that were inattentive or inconsistent. These checks should be randomly inserted in the study, and should appear the same as other trials; otherwise, participants may figure out which trials are the checks. Any data exclusion criteria must be reported for replicability. For studies where answers are subjective, such as aesthetic preferences, another option is to duplicate some trials and verify that the participant's answers are self-consistent. However, there should be enough checks to rule out the possibility of good-faith participants answering randomly on examples that are genuinely difficult to judge.

Choose wording carefully. The way the task or question is phrased can also significantly affect the results [33]. Wording should reflect the high-level goals, e.g., "which image contains fewer artifacts?" instead of "which image contains fewer color defects in the facial region?" Conversely, imprecise task wording leaves too much to interpretation, e.g., "which image is better?" may be understood as "which is more aesthetically-pleasing?" where the intention might have been to evaluate "which is more realistic?" Pilot studies with different wordings can help reveal potential biases due to question wording.

Sample participants representatively. It is important to recruit participants representative of the target user profiles who will ultimately use and/or be impacted by your technology. For instance, if building software for animators, recruit animators to provide feedback. It is also important to recruit inclusively, including, for example, people with disabilities. A common approach of convenience sampling or friend sourcing, whereby a researcher leverages an easily-accessible population, often on a volunteer basis, can lead to biased results that won't replicate.

Augment ratings with qualitative feedback. Running a survey to get quantitative scores does not provide any information about *why* participants gave certain scores. We recommend follow-up interviews or questionnaires to understand *why* participants gave the responses they did. These may help debug the study, but they may also reveal unexpected facets of the output that influenced users' ratings. Was the participant "very unsatified" with the output because it was unrealistic, not aesthetic, biased, or for some other reason? Without this information, the researcher may end up optimizing features of the algorithm that don't address, or may even amplify, the underlying user problems (e.g., focusing on improving realism at the expense of less variable or more biased results).

3.2 Interface Evaluation

Examples of interfaces that appear in computer graphics papers include creativity support tools that help users create graphic designs (e.g., banners, logos, illustrations) or interfaces that help users sculpt 3D meshes. For generative AI tools entering mainstream use, the user interface becomes key to how users interact with the model and specify their needs, expectations, and desired outputs. When the interface itself is a research contribution instead of, or in addition to, the outputs that it generates, then an *interface evaluation* should be performed. In UXR terms, this is an *evaluative* approach which can yield both *behavioral* and *attitudinal* data. Interface evaluation can be used to collect *qualitative* feedback and *quantitative* data (e.g., number and location of clicks, time-to-complete, eye gaze, etc.). Ideally, both *qualitative* and *quantitative* approaches are combined (i.e., *mixed methods*) to produce a robust understanding of how the interface is used.

In a typical study, participants may be recruited to test the functionalities of the tool while generating sample outputs. Methodologically, this may take the form of a usability study or a concept test. A *concept test* is used to examine initial interface design approaches; it typically occurs earlier in the research & development timeline (Figure 3.1). Concept testing involves participants walking through an early interface design concept: not every button needs to be perfected nor every path

through the experience mapped out. Instead, a concept provides the general gist of an experience, which can be shared with users to glean both *behavioral* data (collected through observation of participant interaction with the prototype) and *attitudinal* data (collected through questioning users about their experience with the prototype). A *usability study*, on the other hand, is used to glean concrete, low-level insights into how well users are (or are not) able to use the tool. A usability study is typically performed after an interface has been designed and implemented, at the end of the research & development timeline. Some quantitative metrics might be captured, like the time or number of clicks/selections required to complete a sample task with this tool (e.g., Jones *et al.* [50]). Qualitative user feedback might also be recorded. Occasionally, a comparison is made to a related tool or an existing workflow in absence of an alternative tool. Interface evaluation may reveal issues with users' interaction patterns, requiring adjustments to the interface design. This evaluation is most useful if the researchers can act on the insights gathered to improve the tool.

Interface evaluations published in the graphics literature almost exclusively report quantitative measurements, e.g., [15]–[17], [98], [105], although occasionally subjective impressions are also gathered from participants or evaluators, e.g., [50], [78]. In the HCI literature, it is more common to see graphics interfaces evaluated with qualitative techniques [42]. Recent contributions on generative AI include interfaces for qualitative evaluation [5]. Both quantitative and qualitative techniques are beneficial for evaluating interfaces, and we recommend gathering both qualitative and quantitative data when possible.

Occasionally, an interface is used to evaluate the effectiveness of an underlying model or algorithm, rather than the interface itself. The goal in these cases may be to compare different approaches using the same interface, and report metrics like sample task completion time, diversity of generated outputs, etc. (e.g. Chaudhuri *et al.* [16]). Most of the same study recommendations apply in these cases, with the exception that the learnings from the study are not used to improve the interface.

3.2.1 Study Recommendations

Consider a tool to help novice users ideate on icon designs. To evaluate the success of the tool, it is important to understand novice users' use of the tool, whether it is indeed usable by novices, and whether it helps them to create icon designs.

Recruit study participants from the target populations. Participants from all the target user groups (e.g., novice, intermediate, advanced users) should be invited. A *screener survey* that queries potential participants' traits and demographics may be useful to determine whether they fit within a target user group. Participants should be recruited in an ecologically valid manner (i.e., in a way that is reflective of real-world scenarios). This includes finding people who would otherwise naturally come across the tool if it existed, rather than relying on friends, family, or co-workers, who may be biased. Each participant should be scheduled for a one-on-one session in which they will be given a task to accomplish with a working prototype of the system.

Understand the participant's context. At the start of the session, participants should be asked questions about the context in which they might use such a system, understanding their goals, motivations, processes, and scenarios in which the system may be used eventually. Ideally, the task and context will be as realistic as possible.

Capture first impressions. After the broad introductory questions in an interview, participants can be provided with the task and prototype. Gathering first impressions on the task/how to navigate the prototype, after providing minimal to no guidance, can provide valuable insights about users' natural understanding of the tool being studied. This approach may reveal participant expectations, mental models, and potentially, new UX opportunities.

Give participants control. Placing participants in full control of the prototype, without providing task guidance or next steps, can reveal how users naturally progress through the task using the interface. This approach may address questions like: do participants get stuck? Do they know where to go next or how to operate the tool settings? The goal is to learn how participants would approach the task using your tool if you were not available to guide them.

Provide guidance if appropriate. Sometimes the research goal requires guiding participants to a specific part of a design or prototype where feedback is most desired. For example, if working with a low fidelity prototype, it may not be informative to have participants freely explore the interface. It might be reasonable to assume that outside a research context, help documentation or an onboarding tutorial may be available to a user before reaching the main interface. In these cases, comparable guidance can be provided.

Test a workflow. If the research objective involves testing a hypothesized workflow, the participant can be guided through the workflow, carefully assessing user expectations and needs at each step. For example, if a participant gets stuck on a part of the interface, before supplying guidance, they can be asked "what do you expect to do next?" If they have reached an impasse and have no further feedback, they can be given hints about what to do next, and asked to comment on the action (e.g., was it easy or difficult, and why?).

Ask lots of questions. Instead of directing or guiding the process, the researcher can ask questions—e.g., why the participant chose a certain tool, what they'd expect to do next, how they expect to do that, where they might look to find what they need, if the prototype is working as expected, etc. Here, it will be especially important to avoid leading questions, as they may guide the participant to make unrealistic choices. For instance, a leading question may direct the participant to a menu that would be otherwise hidden, or it may cause the participant to take a step that they would otherwise omit. This may result in overly optimistic results. After the participant completes the task, it is important to ask for feedback on the overall experience, which may be surprisingly different from the feedback provided during the task itself. It's helpful to ask what the participant would change about the experience, as well as if, when, and how they would use the system.

Collect data for data-based claims. There is a plethora of data that may be collected throughout the interface evaluation study. Participants' attitudes towards using the prototype for the given task can be collected as qualitative data (constructive feedback). How the participant interacts with the prototype can also serve as qualitative behavioral data, e.g., where do they indicate confusion by pausing or

asking questions? Quantitative behavioral data can be captured by measuring participants' number of clicks, time-to-complete, eye gaze, etc. This data can be aggregated across participants to deduce trends; additionally, similar measurements can be collected from participants' interactions with existing or competitor systems to drive statistical comparisons. Aside from fueling data-based claims about the efficacy of the designed system, interface evaluations also provide concrete feedback to improve the user experience of the system; in the case of a concept test, this feedback is harvested early in the design process to ensure that the ultimate interface is usable and effective.

3.3 Foundational Research

Compared to evaluations of interfaces and outputs, *foundational research* occurs prior to the start of development to learn more about a problem space, and help identify potential audiences and use cases. These insights provide direction to design, research, and development.

In UXR, the most well-known type of foundational research is *needfinding*: the process of engaging with potential users to uncover opportunities, through interviews and by observing their behaviors in context. It is a *qualitative*, *generative* approach that can yield both *behavioral* and *attitudinal* data. Needfinding typically occurs at the outset of a project, informing what the team ultimately builds, sometimes in ways that would never have occurred to the researchers or developers otherwise. Needfinding plays a crucial role in the iterative design process for commercial product development, which has been shown to lead to more successful products [11], [88], [94], [108]. We believe these benefits will apply to academic graphics and vision research that aims to benefit real users. Needfinding is widely used in the HCI literature. There is also value in exploratory research based on the researcher's own intuitions, but with caution to be wary of the false-consensus effect [12], [89], i.e., overly generalizing from one's own experience to others.

Indeed, we have often witnessed computer graphics or vision researchers attempting to get their research adopted by industry practitioners, only to find that the research does not address the target users' needs. Researchers who do not perform needfinding at the outset may

be surprised to find that users have no need for or interest in the tool they have spent months or years developing. Such tools may perform poorly in evaluation studies, as users may find that the technology produces unhelpful, irrelevant, or unexpected results. Needfinding can help build technologies that will perform well when evaluated by real users, and may be more likely to find a home in commercial tools.

Foundational research can also gather *generative* insights such as design principles and expert workflows, from experts or "lead" users in a field [37]. These insights can be used for developing tools for either experts or novices. For experts, foundational research can identify stages of existing workflows that can be improved with new algorithms or interactions. For novices, design principles gleaned from experts can be embedded into software tools or can provide automated guidance. An example is the survey by Agrawala *et al.* [1] of effective visualization tools designed by extracting principles from expert-drawn illustrations.

3.3.1   Examples

Since needfinding is less common in the graphics and vision literature, we describe two research projects that benefited from it, based on interviews with the paper authors. In each case, needfinding provided important insights that guided the direction of the research; the authors believe that needfinding made a significant difference in their ability to produce useful systems. Our first example is in digital painting. Motivated by the limitations of conventional color pickers for digital painting, Shugrina *et al.* [99] sought to create a better color picker and mixer. The researchers conducted a needfinding study to examine how artists used paint palettes in their daily work, to discover the properties that a tool should have to support these workflows. Based on these studies, the authors designed their Playful Palette tool to combine the physicality of oil and watercolor palettes with digital affordances. The authors then evaluated their tool with another group of artists, demonstrating that it was useful and effective for supporting their creative tasks. They told us that "with a questionnaire, one can struggle to get insightful answers from participants, but observing them do something can uncover their processes, including possible struggles and

inefficiencies". By starting with an understanding of users' actual needs, the authors were able to build a system that met those needs and achieved positive results in the evaluation.

Our second example comes from Zhao *et al.* [112], who built a tool for novices to partially automate icon design. The authors started by studying professional icon designers to gain insights about what aspects of their workflows could be automated. They collected attitudinal data through one-on-one interviews and behavioral data through user observation sessions. These studies revealed that professional designers often collect references from various websites, and from these reference images recombine visual features into the final icon design, later refined using vector editing operations. Informed by these observations, the researchers built the Iconate tool to automatically suggest recombined icons for various user queries. Usability tests with novices provided qualitative data that users liked the icons they generated with the help of the tool, and quantitative data supported that they were faster than when using other conventional tools. Evaluative interviews with professionals demonstrated that they too found the tool useful. The difference between the two target user groups was that professionals only wanted the tool for prototyping ideas, but preferred to complete their designs in external vector editing tools, while novices were keen to use the tool for generating the final icon designs. Over the course of the initial and final interviews, the tool both improved to better fit the needs of the target users, and the paper was more strongly motivated with user needs in mind.

3.3.2 Study Recommendations

Foundational research should come at the start of a project, and will be most beneficial before any technology has been created. Needfinding, the most common type of foundational research, is composed of four core stages [76]. In the first stage, *framing and preparation*, one determines fundamentals such as the research goals and the target, and performs a literature review. In the second stage, *observation and recording*, the researcher immerses themselves in the target audience's context to observe their behavior. The researcher must consider how their presence

in the audience's context might affect behavior, and take measures to minimize disruption to individuals' natural behavior. Third, in *asking and recording*, the researcher interviews the target end users in their own environment, directly recording their words. Last, during *interpretation and reframing*, data is framed in terms of the problems the users need to be solved to improve their situation.

Identify the target users. Identifying an audience or target user group for one's technology at the outset of the project will naturally focus the development and lead to better design. A tool that is targeted too broadly may not end up satisfying the needs of any group. For instance, a graphic design tool built for novice users will likely have a different set of attributes than one for professional graphic designers. Professionals might not want a one-click/one-touch solution, but to have much more control over the creative process, compared to novices who may benefit from fewer options and more automation (e.g., Zhao *et al.* [112]). Similarly, in designing a tool for photographers, there may be important differences in user needs and workflows, for professionals versus amateurs, or portrait versus landscape photographers. Needfinding can help discover these differences and identify the best target user segment for evaluative research.

Recruit users in the wild. Once a target user group has been selected for study, participants should be recruited in naturalistic settings, i.e., where they naturally congregate. For example, there may be a relevant active online message board or social media community that can be leveraged. Even meeting one member of the community may lead to *snowball sampling*, in which relevant users offer connections to similar individuals in their network.

Observe users in their natural environments. In an observation study, there is limited interaction with the users; one simply observes users going about their standard processes, while noting any apparent "pain points" or "aha" moments. These insights help inform the technology that will be built—it should close a gap, address a pain point, or enhance a current process. Prior to observation, consent may need to be collected, considering institutional guidelines.

Seek to understand the users' contexts and needs. One-on-one interviews with target users can be conducted in lieu of or

in addition to an observation session. In these interviews, researchers should start off as broadly as possible, seeking to understand users and their contexts. Examples include mapping out their current workflows, and probing their motivations and decision making processes. If the goal is to understand a specific workflow, users can be asked to walk through it step-by-step. Screensharing can be used with remote users. It is useful to combine interviews and observational sessions (a methodology called *contextual inquiry*) to best understand target users from both *behavioral* and *attitudinal* perspectives. This may also take the form of an initial interview, followed by an observational session. Asking participants questions about their process as they perform it in real-time is a convenient way to probe certain areas of interest.

Minimize bias during your interviewing. The way in which a researcher structures an interview and asks questions can bias participant responses. First, it is important to make the participant feel as comfortable as possible by clarifying that there are no right or wrong answers. Making space for silence instead of finishing the participant's sentences, and avoiding providing feedback throughout the interview (even saying "I agree") will help minimize bias in their answers. Leading questions should be avoided (as discussed in Sections 3.1, 3.2). Other types of interview questions to avoid are future behavior predictions (e.g. "What do you think your process will look like in the future? Would you ever use a tool like this?"); numerous psychology papers (for reviews see Armor *et al.* [3] and Dunning [31]) have taught us that humans can not accurately predict their future behaviors. Double-barrelled questions (e.g., "What do you think about performance, speed, and accuracy? How do you decide which tool to use on a given day?") should be broken up, to ask one question at a time (the example above could be four separate questions).

Write a discussion guide. One way to minimize bias is to prepare and review the questions in advance. A *discussion guide* is a set of user-facing questions, distinct from the study's internal research questions, formulated specifically to learn relevant information from participants. The first questions in the guide should include some "warm-up" questions, not specifically tied to the goals of the research (e.g., "Where are you calling in from today?", "What do you like to do for fun?"), to

help participants become more comfortable with the interview. These questions may also provide some insight into the participants' context and what is important to them. The discussion guide questions can then get increasingly narrower in scope, including questions about workflows, pain points, and finally, honing in on participant reactions to the presented solution. The discussion guide includes a list of questions to choose from to provide structure, and ensure the research goals are addressed, but is just a guide. We recommend taking a *semi-structured* interview approach, which means the researcher can follow-up on certain answers or ask new questions as they emerge.

Consider your own assumptions and biases. The researcher's perspective is influenced by their role in conducting the research, their own experiences, and assumptions [22]. *Reflexivity* refers to examining one's own beliefs, judgements, and practices during the research process, and how these factors can impact the research [34]. Limitations identified should be noted by the researcher throughout the research process, and considered during data analysis. This practice may shape how we approach reporting results, including contextualizing the research through a particular lens, or reporting biases as limitations.

4

The Broader Landscape of User Research Methodologies

There are many ways to gain important insights from human participants throughout the research & development lifecycle, from conceptualizing a tool to iterating on the final designs. Depending on the research goals and the stage of the research, different user research methodologies may be appropriate. In this section, we provide an overview of this broad landscape of methodologies (Figure 4.1), with pointers to further reading.

The type of data one will need to address their objectives, questions, and decisions can guide the choice of methodologies. For instance, attitudinal data (i.e., what users *think*), can be collected with a survey; while behavioral data (i.e., what users *do*) can be recorded during observational sessions.

Considering the current state of the project will further narrow down the relevant methodologies. Foundational research is useful during early ideation, while output evaluations are conducted with a fully-built model. Generative methods are conducted in the early, discovery stages; they can help define a target user to guide research or product development. Conversely, if the target user has already been established, and the product concept validated, a usability study can evaluate the

implementation. Evaluative methods generally apply when assessing functionality, from an early-stage concept to a fully-built product, and are typically used in the later stages of a project.

4.1   Other Study Designs

Table 4.1 provides a working definition for the user research methods we posit are most applicable to graphics and vision researchers (i.e., in the darker color in Figure 4.1), and which we already discussed in Section 3. In the lighter color are other frequently used UX research methods that are currently underutilized by vision and graphics researchers. For instance, *focus groups* allow several participants to be simultaneously present, engaged in group discussion facilitated by a moderator. Focus groups are well-suited if a researcher has already identified core themes they want to further explore with participants. Such themes may emerge from earlier observational studies or one-on-one interviews. *Diary studies* are particularly useful for reviewing a participant's experience over time, and tracking behaviors (e.g., software usage) over a multi-week period. Participants are typically supplied with pre-defined prompts or questions to answer throughout the study. *Participatory design* is a process by which researchers and participants co-create an interface, not to be used as-is, but to reveal hidden user needs and considerations. *Eye tracking* can be used to track and show where a participant is looking, typically to understand how the participants process a given user interface.

4.2   Alternatives to Formal User Studies

While the previous section lists formal user research methods, there are ways to capture usability issues *without* running a user study, e.g., with a usability inspection method [67], such as a cognitive walkthrough [84] or heuristic evaluation [68]. In a *cognitive walkthrough*, the researcher puts themselves in the shoes of a potential user (e.g., see Section 5 in [5]). The researcher attempts to use the tool, and as they proceed, they evaluate the usability of various potential paths through the user experience. In this process, they consider how users will access various features, where and when they will be able to proceed, potential areas
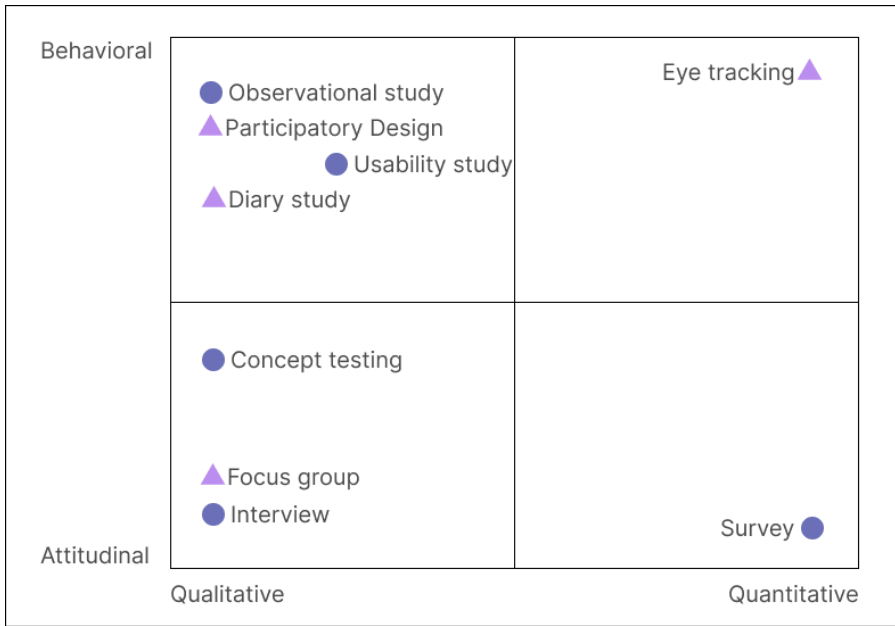
Figure 4.1: Dark color circles are the methods discussed at greater length in this paper, and lighter color triangles are other relevant methods that are mentioned in passing. The user research landscape is larger than represented here, with more study designs possible, but these particular examples were curated with the vision and graphics research audience in mind (adapted from [86]).

of confusion, etc. A *heuristic evaluation*, while similar, is conducted by an expert and relies on pre-defined quantitative metrics. The heuristic evaluator also steps through the user experience, but they rate the experience on a predefined set of established heuristics (e.g., Nielsen *et al.* [68]) as they proceed. This is particularly useful for highlighting usability issues and areas of improvement in the user journey. Such evaluations can replace a formal user study under limited resources or be used as a precursor to further user testing. Learnings from such evaluations can be used throughout research & development to refine tools.

Table 4.1: Definitions of common methodologies in user experience research. Further reading on user research methods include [24], [37], [93].

Methodology	Definition
Observational study	Observation of participants in their naturalistic settings, without any researcher intervention or manipulation of variables. Also referred to as "field research". *Further reading: Nielsen [66]*
Interview	One-on-one conversation between the participant and researcher, in which the researcher asks qualitative questions to the participant to collect attitudinal data. Could be used as building blocks in other methods (e.g., concept testing). *Further reading: Arsel [4] and Portigal [81]*
Concept testing	Using designs, often of low fidelity, to test concepts, conducted early in the research and development process, serving as guidance for broader project direction. Provides information about whether you are solving the problem you hypothesize you are solving. *Further reading: Sharon [97]*
Usability study	A method to evaluate the way in which a participant interacts with a user-facing interface; typically, this method examines the ease-of-use, understandability, performance, and user experience of an interface. This method is typically conducted late in the research and development process, after the concept direction has been validated earlier in the process, via concept testing. *Further reading: Albert et al. [2] and Rubin et al. [90]*
Survey	A method to collect self-reported user data at scale, addressing questions like "what percentage of users do x or y?" *Further reading: Rea et al. [82]*

4.3 Combining Methodologies

Synthesizing the results of multiple study methodologies, via a *mixed methods* approach, is likely to provide richer insights. Including both behavioral and attitudinal, as well as qualitative and quantitative data, will provide the most holistic picture of the user.

It is common for researchers to over-rely on surveys as a quick way to get quantitative data. There is a widespread tendency to trust quantitative data over qualitative data [60], which may be exacerbated by an unfamiliarity with qualitative techniques, a point we hope to address with this work. Simply adding open-ended questions into a quantitative survey does not harness the qualitative data we are referring to; most participants will not spend the time writing out responses to open-ended questions the same way they might extrapolate on their needs in a 1:1 interview.

Another natural outcome of formative user studies could be the criteria that users may judge a tool or output by. Interviews and focus groups can give researchers insights on features of a tool or output that are most important or relevant to the users and their workflows, which can then be converted into criteria or metrics for evaluation.

Consider the example of measuring creativity or evaluating creative outputs. A single quantitative measure will not capture the complexity of what it means for an output to be more or less creative. Past research on creativity-related technology has utilized a bevy of disparate evaluation strategies. One oft-cited study [38] concluded that more diverse alternatives will afford a more creative outcome, leading many researchers to prioritize quantity as a proxy for creativity. Other researchers took this a step further, arguing that more efficient tools will enable the creation of more design outputs [27], thereby fostering creativity. This approach prizes speed over quality by measuring time-to-complete as a proxy for creativity. Yet other researchers have hired *design experts*, a tricky pool to define and recruit, to arbitrarily rate the quality of creative outputs [63]. None of these approaches single-handedly capture creativity, necessitating a mixed methods approach [25]. As an example, Herman and Hwang [44] propose a mixed methods framework for measuring creativity, in which qualitative, quantitative, behavioral, and

visual data are triangulated to formulate a user-centered evaluation of creative tools.

5

Reporting and Conducting Ethical Research

The danger of treating user studies as a second-class citizen is that important methodological details may be left out and terminology may be used incorrectly, impeding future reproducibility efforts. Further, providing too few details hinders the ability to evaluate whether the conclusions made based on those user studies are valid and reproducible. In this section we advocate for careful reporting of methods and conclusions, and for ethical procedures to be followed when working with participants.

5.1 Reporting Requirements

In computer science paper writing and reviewing, it is common to pay laser-focused attention to the description and reproducibility of algorithms. On the other hand, user study descriptions are treated as an afterthought. We argue that it's not enough to perform a study carefully, it must also be reported thoroughly, so that the reader can understand what was done, what the results show, and how they can be reproduced. Authors should also take care in how they present and interpret their own results.

Report methodology reproducibly. Descriptions of user studies should include enough details for reproducibility, including how participants were recruited, a description of the participants (e.g., occupation, experience level in relevant domain), stimuli used (e.g., interface prototype, number and type of images), procedural details (e.g., interview length, structure, examples of questions asked) [95]. Since the exact question wording can greatly affect participant responses (Section 3), this information should be made available, for instance in supplementary material and/or in online repositories.

Use precise terminology. Computer graphics and vision papers typically use imprecise terminology around output evaluations, beginning with the term *user*. The term *user study* originated as a description of an evaluation with prospective users of a real system, for example, an evaluation of word processing software with people who might want to write documents. The crowdworkers in computer graphics and vision studies would more accurately be described as *participants*, *raters*, *evaluators*, or *labelers*. Likewise, the term *user study* as a blanket description for all studies lacks precision, leaving the actual type of study opaque. Referring to a study's design by a precise name can facilitate both study planning and reproducibility, by directing researchers to relevant and related work. Table 1 lists a set of methodologies that all fall under the umbrella term *user studies*, and includes references for further reading on best practices. The present blanket use of the term *user study* in graphics and vision papers gives the misleading impression that all user studies are basically the same. When possible, we recommend using more specific terminology. For example, for the common evaluation of outputs done with surveys, we recommend "human rating study".

Interpret results carefully. One danger occurs when authors overgeneralize from an experiment with a limited set of stimuli and narrow participant pool to very broad and general conclusions. For example, one study in a psychology journal [36] concluded that "individuals are unable to accurately identify AI-generated artwork," from comparisons of GAN-generated images against a few existing artworks, made by crowdworkers viewing low-resolution, out-of-context imagery. This paper does not report details about which specific image sets were collected or used, making the claims hard, if not impossible, to verify and reproduce.

More worrisome is that the popular press reported these results with the misleading claims that AIs can make art as well as humans.

5.2 Conducting Research Ethically

The previous sections outlined high-level principles for user studies. There are also core ethical aspects that should be part of every user study, including obtaining consent, compensating fairly, and conducting safe and inclusive research.

How to get consent? Different institutions have different requirements (e.g., Institutional Review Board approval) for dealing with human participants. It is important to follow this guidance, procuring informed consent if necessary. If an institution does not require an approval process, it is still highly recommended to receive proper consent from participants before they participate in studies. Participants should be informed of the intent of the study, what data will be collected, how it will be used and stored, and any foreseeable harms or risks. It is good practice to also verbally communicate this information to study participants and ask for necessary permissions (e.g., to record, take pictures, etc.). This helps protect participants and make them feel more comfortable. Getting consent is part of a broader practice of conducting ethical research with human participants (e.g., Jhangiani *et al.* [48], pp. 46-66; for more information on informed consent). Further, all *personally identifiable information (PII)* collected from participants should be anonymized (unless the study design requires otherwise) and stored securely, following local data privacy rules.

How to compensate? One of the most important considerations for unbiased and ethical research is fair compensation of participants. Researchers should be aware of the minimum pay rates in the regions where the research is run, e.g., certain U.S. states have a higher rate than the national minimum. Even so, paying at the minimum might not encourage very active or eager participation. With the proliferation of remote work and schooling, we have seen increasing difficulties in motivating study participation. Under limited resources, researchers can consider internally motivating participants (see also Cialdini [20]), for instance by gamifying the studies and providing participants with actionable feedback or personalized results [83].

How to do this all safely and inclusively? The COVID-19 pandemic has increased the amount of user research carried out virtually via online video conferencing software. Remote user research aids in disability and geographical inclusivity, since it is possible to include users from different regions and those who may not easily be able to travel to an in-person meeting. Virtual sessions allow participants to share their screen as they walk through previous projects and/or the proposed prototype, so the researcher can see exactly how they're attempting to interact with the interface. Many videoconferencing options also enable efficient recording and transcription services. When conducting international research, one should be sensitive of timezones, compensation using local currency, and having translators on-hand if necessary. Bloom *et al.* [9] includes a full guide on running inclusive research with people with disabilities, who may require accommodations.

6

Limitations and Problems of User Studies

In graphics and vision communities, it is difficult to publish papers that do not include a table of scores that show better numbers for the proposed method compared to previous methods. But, for user evaluations, whether or not any of these numbers are meaningful is rarely scrutinized. When the reviewing culture makes user studies a requirement for publication, without simultaneously providing standards for how to run them properly—or whether to run them at all—it can harm research progress. We make recommendations to help researchers, paper reviewers and conference chairs combat these effects.

6.1 Replicability and Reproducibility

When we perform an experiment, such as testing the output of an algorithm, the results should be meaningful. Computer graphics and AI reviewing emphasize *reproducibility* of the algorithm itself, e.g., Bonneel *et al.* [10] and Pineau *et al.* [79]. *Reproducibility* means that the exact same experiment can be performed on the same data, e.g., when the authors release their data and code ([64], p. 43). The results should also be generalizable, so the gold standard for experimental findings in mainstream science is *replicability* ([64], p. 43 based on Barba [6]).

Findings are replicated when another researcher performs the same test on different data and/or different participants, and obtains the same result. For example, if another researcher were to compare your published algorithm to your baselines, on a new collection of users evaluating the results, would your algorithm perform just as well as when you tested it? If an experiment cannot be replicated, we generally treat the original findings as invalid. Note that definitions of reproducible/replicable are sometimes swapped [6], [80].

The Replication Crisis in mainstream science provides a cautionary tale [74]. Following these developments, many authors have raised replicability concerns for computer science [21], [77], specifically within machine learning [41], [79] and HCI [39], [40], [46], [111]. Many of the causes cited for these replication crises are relevant to computer graphics and vision, e.g., researchers can manipulate results–even unintentionally–such as by only reporting results from successful trials [100].

As a case study, Wang *et al.* [110] performed several MTurk user studies testing versions of the same hypothesis, and got different results from different tests. While these studies were meant as exploratory, they illustrate how different studies may reach different conclusions on the same question.

6.2   Over-emphasis on User Studies Can Hinder Progress

Quantitative evaluations play an important role in computer science research to directly compare methods, and judge whether a new idea translates into improvements. However, numerical evaluations can be flawed for many reasons, and an overemphasis on imperfect metrics can distort research, as discussed in many prior contexts [8], [28], [38], [103].

A good case study is an ICLR 2022 paper for which the reviews are available online [47]. Two reviewers gave very negative scores due, in part, to a lack of user studies. The paper was eventually accepted, accompanied by a summary chastizing the reviewers for using "user studies" as an excuse for poor reviewing, and accusing them of gatekeeping. The final decision noted that the submission described a software library that had been deployed for years, with thousands of users (information that was not revealed to the reviewers for anonymous review).

Would the paper—which describes a highly impactful system—have been rejected if the committee had not had this information? And, had the authors gone through the extra effort of contriving and performing a user study, would it have been meaningful, and would it have been enough to convince the reviewers?

Too-early quantification discourages innovation. This emphasis on studies becomes harmful to the field when it discourages publication of innovative ideas that cannot easily be evaluated. In computer graphics research, we often seek to design new ways for users to create and design images, 3D, video, artwork, etc. For example, as pointed out by Greenberg *et al.* [38], in a searing critique of the over-emphasis on evaluation in HCI research, radical new interaction designs naturally will perform worse in many kinds of evaluations than polished mature systems. User research for early-stage designs (i.e., foundational research) must naturally be different from the numerical evaluation one might perform on a mature system.

Reviewers and editors are responsible. We have observed reviewers and editors impose onerous evaluation requirements on papers, without thought for whether such evaluations are meaningful or necessary. We have also observed authors and reviewers use crowdsourced evaluations as a crutch to avoid making hard decisions. Reviewer comments like "I can't tell if the images are better, maybe a user study would help" are potentially harmful, encouraging authors to perform extra work that will not improve a lackluster paper.

Recommendations. We recommend that paper reviewers and area chairs take a critical eye toward evaluation: when requesting an additional user evaluation for a paper, to consider whether the evaluation would be meaningful or necessary for the paper. Some of these problems can become systematic: the standards of what reviewers expect in a paper arise organically, as a result of authors' own experience in previous papers, and what previous reviewers expected of them. Eventually these trends become common expectations, requiring a broader conversation in the field, and the efforts of program chairs to remind reviewers and area chairs to take the evidence and ideas in the paper as a whole, and not just look for the table of numbers.

6.3   Bias and Representation in Study Populations

Noble [69] describes numerous examples where user research in product development led to biased systems because of failures to recruit diverse participants. Noble demonstrates that developers' biases were not revealed in user studies because the studies performed included only similarly biased, white, cis-gender men as participants. Noble reveals deep-seated (and often unintentional) discrimination that has been baked into Google Image Search, Google PageRank, Airbnb rental offerings, ArtStor's metadata, among others, resulting in biased outputs. In the context of graphics and vision work, it's worth considering whether a system designed to reconstruct a user's hair would, for instance, work for different types of hair [51]. Further, Noble *et al.* [70] discussed how algorithmic outputs are often evaluated according to pre-existing social norms and biases that include racism and sexism.

Recommendations. Unfortunately, we're not aware of any quick and easy solutions for approaching maximally inclusive research participant pools. There remain many open research questions, including which range of identities should be included in each study. A team might consider recruiting participants that exhibit a range of gender identities, ethnicities, sexual orientations, international locations and physical and mental abilities, depending on the technology being tested. For instance, when producing a tool that pertains to colors, it may be useful to include the perspective of someone who is colorblind. When constructing a novel interface, it will be helpful to include someone who uses a screenreader to determine its usability. For software that pertains to images of the human body (e.g., human hair or poses), it will be helpful to include a range of body types and ethnicities. For more tips on performing user research with participants with disabilities, see Herman *et al.* [44].

7

Conclusion

User studies have occupied a contradictory place in computer graphics and vision research. On one hand, user studies are often considered a necessary step in the evaluation of research results. On the other hand, authors and reviewers often treat user studies as an afterthought, not leaving sufficient time for the careful planning and reporting that replicable research requires. Furthermore, penalizing work that does not contain user evaluation has the unintended consequence of incentivizing hastily done, poorly executed user research. A maxim to keep in mind is that **"bad user research leads to bad outcomes"**, and such research will continue if reviewers continue to ask for it. Our call to action for the computer graphics and vision communities is to put more thought into the requests for, design, and execution of user studies.

On the other hand, when treated as a first-class citizen, commensurate with the algorithmic contributions of a paper, user studies can provide valuable insights to help researchers iterate on and improve the design of their algorithms or interfaces. For instance, foundational research, including interviews and needfinding techniques, invoked at the start of the research and development process, offers the best ways to learn about users' wants and needs, even elucidating unexpected

potential research directions. For interfaces designed to support user workflows such as creative tasks, concept testing earlier prototypes or design concepts can inform future iterations, and usability studies can be invoked to evaluate later-stage prototypes. For evaluating algorithm outputs such as images or videos, qualitative feedback gathered alongside quantitative survey scores can help researchers understand how participants rate outputs, and which features are most important to them (which may be different from which features the researchers chose to optimize). In all cases, we argue that user evaluations are more useful if the researchers can act on the insights gathered to improve the algorithm or tool, rather than viewing user studies as the step at the end of the research & development road.

While crowdwork platforms have made certain types of user evaluation - particularly surveys - easier and faster to complete, they represent a narrow slither of the user research landscape. As an introduction to this broader landscape, and with an eye to what is most relevant to vision and graphics research, we covered a number of methodologies from user experience (UX), human-computer interaction (HCI), and related fields that can be invoked at various time points throughout the research & development lifecycle. We categorized user research into foundational research, interface evaluation, and output evaluation, and used the UX axes qualitative/quantitative, attitudinal/behavioral, and generative/evaluative when describing different methodologies. This additional precision in terminology helps elucidate the types of human data being captured in the evaluation. For each of the categories of user research, we provided actionable guidelines and best practices, while referencing relevant resources for further reading. With this guidance in hand, we hope to contribute to improved quality standards for user studies in computer vision and graphics research.

Acknowledgements

References

[1] M. Agrawala, W. Li, and F. Berthouzoz, "Design principles for visual communication," *Commun. ACM*, vol. 54, no. 4, Apr. 2011, pp. 60–69. DOI: 10.1145/1924421.1924439.

[2] W. Albert and T. Tullis, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

[3] D. A. Armor and S. E. Taylor, "Situated optimism: Specific outcome expectancies and self-regulation," *Advances in experimental social psychology*, vol. 30, 1998, pp. 309–379.

[4] Z. Arsel, "Asking questions with reflexive focus: A tutorial on designing and conducting interviews," *Journal of Consumer Research*, vol. 44, no. 4, 2017, pp. 939–948.

[5] Y. Assogba, A. Pearce, and M. Elliott, "Large scale qualitative evaluation of generative image model outputs," *arXiv preprint arXiv:2301.04518*, 2023.

[6] L. A. Barba, *Terminologies for reproducible research*, 2018.

[7] T. L. Berg and A. C. Berg, "Finding iconic images," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, pp. 1–8, 2009.

[8] M. Biagioli, "Watch out for cheats in citation game," *Nature News*, vol. 535, no. 7611, 2016, p. 201.

[9] M. Bloom, E. Ellis, N. Era, L. Herman, F. Kazerooni, L. Meckler, and E. C. Siegel, *A practical guide to inclusive research*, 2021. URL: https://medium.com/inclusive-research.

[10] N. Bonneel, D. Coeurjolly, J. Digne, and N. Mellado, "Code replicability in computer graphics," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, 2020, pp. 93–1.

[11] B. Brown and R. Hart, *Forrester infographic: The roi of design thinking*, 2019. URL: https : / / www . forrester . com / report/Forrester-Infographic-The-ROI-Of-Design-Thinking/ RES157995.

[12] R. Budiu, *You are not the user: The false-consensus effect*, 2017. URL: https://www.nngroup.com/articles/false-consensus/.

[13] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?," 2016.

[14] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 40–48, 2018.

[15] S. Chaudhuri, E. Kalogerakis, S. Giguere, and T. Funkhouser, "Attribit: Content creation with semantic attributes," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 193–202, 2013.

[16] S. Chaudhuri, E. Kalogerakis, L. Guibas, and V. Koltun, "Probabilistic reasoning for assembly-based 3d modeling," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011. DOI: 10.1145/2010324.1964930.

[17] C.-H. Chiu, Y. Koyama, Y.-C. Lai, T. Igarashi, and Y. Yue, "Human-in-the-loop differential subspace search in high-dimensional latent space," *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020. DOI: 10.1145/3386569.3392409.

[18] M. Chmielewski and S. C. Kucker, "An mturk crisis? shifts in data quality and the impact on study results," *Social Psychological and Personality Science*, vol. 11, no. 4, 2020, pp. 464–473. DOI: 10.1177/1948550619875149.

[19] T. S. Cho, N. Joshi, C. L. Zitnick, S. B. Kang, R. Szeliski, and W. T. Freeman, "A content-aware image prior," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 169–176, 2010.

[20] R. B. Cialdini, *Influence: The psychology of persuasion*, vol. 55. Collins New York, 2007.

[21] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin, "Threats of a replication crisis in empirical computer science," *Communications of the ACM*, vol. 63, no. 8, 2020, pp. 70–79.

[22] L. Cohen, L. Manion, and K. Morrison, *Research methods in education.* routledge, 2002.

[23] F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh, "How well do line drawings depict shape?" In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 28, Aug. 2009.

[24] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications, 2017.

[25] N. Crilly, "Methodological diversity and theoretical integration: Research in design fixation as an example of fixation in research design?" *Design Studies*, vol. 65, 2019, pp. 78–106.

[26] C. Cuskley and J. Sulik, *The burden for high-quality online data collection lies with researchers, not recruitment platforms*, 2022. DOI: 10.31234/osf.io/w7qy9.

[27] S. Davidoff, M. K. Lee, A. K. Dey, and J. Zimmerman, "Rapidly exploring application design through speed dating," in *International Conference on Ubiquitous Computing*, Springer, pp. 429–446, 2007.

[28] M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals, "The benchmark lottery," 2021.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248–255, 2009.

[30] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," *Advances in neural information processing systems*, vol. 28, 2015.

[31] D. Dunning, "Self-image motives and consumer behavior: How sacrosanct self-beliefs sway preferences in the marketplace," *Journal of consumer psychology*, vol. 17, no. 4, 2007, pp. 237–249.

[32] D. G. Elmes, B. H. Natowitz, and H. L. Roediger, *Research Methods in Psychology, 6th Edition.* Pacific Grove, California: Brooks/Cole Publishing Company, 1999.

[33] Z. Epstein, S. Levine, D. G. Rand, and I. Rahwan, "Who gets credit for ai-generated art?" *Iscience*, vol. 23, no. 9, 2020, p. 101 515.

[34] L. Finlay, "Reflexivity: An essential component for all research?" *British Journal of Occupational Therapy*, vol. 61, no. 10, 1998, pp. 453–456.

[35] C. Fosco, V. Casser, A. K. Bedi, P. O'Donovan, A. Hertzmann, and Z. Bylinskii, "Predicting visual importance across graphic design types," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 249–260, 2020.

[36] H. Gangadharbatla, "The role of ai attribution knowledge in the evaluation of artwork," *Empirical Studies of the Arts*, 2021. DOI: 10.1177/0276237421994697.

[37] E. Goodman, M. Kuniavsky, and A. Moed, *Observing the user experience: A practitioner's guide to user research.* Elsevier, 2012.

[38] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 111–120, 2008.

[39] S. Greenberg and H. Thimbleby, "The weak science of human-computer interaction," 1991.

[40] C. Greiffenhagen and S. Reeves, "Is replication important for hci?," 2014.

[41] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, and C. S. Greene, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, 2020, E14–E16.

[42] J. Hartmann, S. DiVerdi, C. Nguyen, and D. Vogel, "View-dependent effects for 360° virtual reality video," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '20, pp. 354–364, Virtual Event, USA: Association for Computing Machinery, 2020. DOI: 10.1145/ 3379337.3415846.

[43] J. Heer and M. Bostock, "Crowdsourcing graphical perception: Using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212, 2010.

[44] L. M. Herman and A. H.-C. Hwang, "In the eye of the beholder: A viewer-defined conception of online visual creativity," *New Media & Society*, 2022, p. 14 614 448 221 089 604. DOI: 10.1177/ 1461444822108960.

[45] A. Hertzmann, "Non-Photorealistic Rendering and the Science of Art," in *Proc. NPAR*, 2010.

[46] K. Hornbæk, S. S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen, "Is once enough? on the extent and content of replications in human-computer interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3523– 3532, 2014.

[47] ICLR, *Reviewer discussion for iclr 2022 submission: "einops: Clear and reliable tensor manipulations with einstein-like notation"*, 2022.

[48] R. S. Jhangiani, I.-C. A. Chiang, C. Cuttler, D. C. Leighton, and M. A. Metz, "From moral principles to ethics codes," *Research Methods in Psychology*, 2020.

[49] D. Jones, *A weird view of human nature skews psychologists' studies*, 2010.

[50] R. K. Jones, D. Charatan, P. Guerrero, N. J. Mitra, and D. Ritchie, "Shapemod: Macro operation discovery for 3d shape programs," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, 2021, pp. 1–16.

[51] T. Kim, H. Rushmeier, J. Dorsey, D. Nowrouzezahrai, R. Syed, W. Jarosz, and A. M. Darke, *Countering racial bias in computer graphics research*, 2021.

[52] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456, 2008.

[53] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide," *Data mining and knowledge discovery*, vol. 18, no. 1, 2009, pp. 140–181.

[54] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman, "Crowdsourcing in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 10, no. 3, 2016.

[55] M. Kuniavsky, *Observing the user experience: a practitioner's guide to user research*. Elsevier, 2003.

[56] L.-J. Li, C. Wang, Y. Lim, D. M. Blei, and L. Fei-Fei, "Building and using a semantivisual image hierarchy," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 3336–3343, 2010.

[57] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 453–468, 2018.

[58] J. Lu, C. Barnes, S. DiVerdi, and A. Finkelstein, "Realbrush: Painting with examples of physical media," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, 2013, pp. 1–12.

[59] J. Lu, F. Yu, A. Finkelstein, and S. DiVerdi, "Helpinghand: Example-based stroke stylization," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, 2012, pp. 1–10.

[60] S. Mau, *The metric society: On the quantification of the social*. John Wiley & Sons, 2019.

[61]   R. McDonnell, M. Breidt, and H. H. Bülthoff, "Render me real? investigating the effect of render style on the perception of animated virtual humans," *ACM Trans. Graph.*, vol. 31, no. 4, Jul. 2012. DOI: 10.1145/2185520.2185587.

[62]   R. McDonnell, M. Larkin, S. Dobbyn, S. Collins, and C. O'Sullivan, "Clone attack! perception of crowd variety," vol. 27, no. 3, Aug. 2008, pp. 1–8. DOI: 10.1145/1360612.1360625.

[63]   P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun, "Interactive furniture layout using interior design guidelines," *ACM transactions on graphics (TOG)*, vol. 30, no. 4, 2011, pp. 1–10.

[64]   National Academies of Sciences Engineering and Medicine, *Reproducibility and replicability in science*. National Academies Press, 2019.

[65]   A. L. Nichols and J. K. Maner, "The good-subject effect: Investigating participant demand characteristics," *The Journal of general psychology*, vol. 135, no. 2, 2008, pp. 151–166.

[66]   J. Nielsen, *Field studies done right: Fast and observational*, 2002. URL: https://www.nngroup.com/articles/field-studies-done-right-fast-and-observational/.

[67]   J. Nielsen, "Usability inspection methods," in *Conference companion on Human factors in computing systems*, pp. 413–414, 1994.

[68]   J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 249–256, 1990.

[69]   S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[70]   S. U. Noble and B. M. Tynes, *The intersectional internet: Race, sex, class, and culture online*. Peter Lang International Academic Publishers, 2016.

[71]   P. O'Donovan, A. Agarwala, and A. Hertzmann, "Color compatibility from large datasets," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–12.

[72] C. O'Sullivan and J. Dingliana, "Collisions and perception," *ACM Trans. Graph.*, vol. 20, no. 3, Jul. 2001, pp. 151–168. DOI: 10.1145/501786.501788.

[73] C. O'Sullivan, J. Dingliana, T. Giang, and M. Kaiser, "Evaluating the visual fidelity of physically based animations," *ACM Transactions on Graphics*, vol. 22, Jul. 2003. DOI: 10.1145/882262.882303.

[74] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

[75] D. Parikh and C. L. Zitnick, "The role of features, algorithms and data in visual recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 2328–2335, 2010.

[76] D. Patnaik and R. Becker, "Needfinding: The why and how of uncovering people's needs," *Design Management Journal (Former Series)*, vol. 10, no. 2, 1999, pp. 37–43.

[77] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, 2011, pp. 1226–1227.

[78] G. Philbrick and C. S. Kaplan, "A primitive for manual hatching," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 2, 2022, pp. 1–17.

[79] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)," *Journal of Machine Learning Research*, vol. 22, 2021.

[80] H. E. Plesser, "Reproducibility vs. replicability: A brief history of a confused terminology," *Frontiers in Neuroinformatics*, vol. 11, 2018, p. 76. DOI: 10.3389/fninf.2017.00076.

[81] S. Portigal, *Interviewing users: how to uncover compelling insights*. Rosenfeld Media, 2013.

[82] L. M. Rea and R. A. Parker, *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons, 2014.

[83] K. Reinecke and K. Z. Gajos, "Labinthewild: Conducting large-scale online experiments with uncompensated samples," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 1364–1378, 2015.

[84]   J. Rieman, M. Franzke, and D. Redmiles, "Usability evaluation with the cognitive walkthrough," in *Conference companion on Human factors in computing systems*, pp. 387–388, 1995.

[85]   D. Ritchie, K. Wang, and Y.-a. Lin, "Fast and flexible indoor scene synthesis via deep convolutional generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6182–6190, 2019.

[86]   C. Rohrer, *When to use which: User-experience research methods*, 2014. URL: https://www.nngroup.com/articles/which-ux-research-methods/.

[87]   M. Rosala, *Rating scales in ux research: Likert or semantic differential?* 2020. URL: https://www.nngroup.com/articles/rating-scales/.

[88]   S. R. Rosenthal and M. Capper, "Ethnographies in the front end: Designing for enhanced customer experiences," *Journal of Product Innovation Management*, vol. 23, no. 3, 2006, pp. 215–237.

[89]   L. Ross, D. Greene, and P. House, "The "false consensus effect": An egocentric bias in social perception and attribution processes," *Journal of experimental social psychology*, vol. 13, no. 3, 1977, pp. 279–301.

[90]   J. Rubin and D. Chisnell, "How to plan, design, and conduct effective tests," *Handbook of usability testing*, vol. 17, no. 2, 2008, p. 348.

[91]   B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, 2008, pp. 157–173.

[92]   J. Sauro, *How many people cheat in online surveys?* 2010. URL: https://measuringu.com/cheat-survey/.

[93]   J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[94]   C. R. Schaffhausen, "Large-scale needfinding methods, quality metrics, and need prioritization in user-centered design," Ph.D. dissertation, University of Minnesota, 2015.

[95] S. Schoch, D. Yang, and Y. Ji, ""this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation," in *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pp. 10–16, 2020.

[96] A. Serrano, B. Chen, C. Wang, M. Piovarči, H.-P. Seidel, P. Didyk, and K. Myszkowski, "The effect of shape and illumination on material perception: Model and applications," *ACM Trans. Graph.*, vol. 40, no. 4, Jul. 2021. DOI: 10.1145/3450626.3459813.

[97] T. Sharon, *Validating product ideas: Through lean user research.* Rosenfeld Media, 2016.

[98] R. Shilkrot, P. Maes, J. A. Paradiso, and A. Zoran, "Augmented airbrush for computer aided painting (cap)," *ACM Trans. Graph.*, vol. 34, no. 2, Mar. 2015. DOI: 10.1145/2699649.

[99] M. Shugrina, J. Lu, and S. Diverdi, "Playful palette: An interactive parametric color mixer for artists," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, 2017, pp. 1–10.

[100] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science*, vol. 22, no. 11, 2011, pp. 1359–1366.

[101] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, IEEE, pp. 1–8, 2008.

[102] M. Spain and P. Perona, "Some objects are more equal than others: Measuring and predicting importance," in *European conference on computer vision*, Springer, pp. 523–536, 2008.

[103] K. O. Stanley and J. Lehman, *Why greatness cannot be planned: The myth of the objective.* Springer, 2015.

[104] J. O. Talton, D. Gibson, L. Yang, P. Hanrahan, and V. Koltun, "Exploratory modeling with collaborative design spaces," *ACM Trans. on Graphics*, vol. 28, no. 5, 2009.

[105] J. O. Talton, D. Gibson, L. Yang, P. Hanrahan, and V. Koltun, "Exploratory modeling with collaborative design spaces," *ACM Trans. Graph.*, vol. 28, no. 5, Dec. 2009, pp. 1–10. DOI: 10.1145/1618452.1618513.

[106] P. Vangorp, J. Laurijssen, and P. Dutré, "The influence of shape on the perception of material reflectance," New York, NY, USA: Association for Computing Machinery, 2007. DOI: 10.1145/1275808.1276473.

[107] S. Vijayanarasimhan and K. Grauman, *Multi-level active prediction of useful image annotations for recognition*. Citeseer, 2008.

[108] E. Von Hippel, "Lead users: A source of novel product concepts," *Management science*, vol. 32, no. 7, 1986, pp. 791–805.

[109] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie, "Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, 2019, pp. 1–15.

[110] X. Wang, Z. Bylinskii, A. Hertzmann, and R. Pepperell, "A computational approach to studying aesthetic judgments of ambiguous artworks," *Psychology of Aesthetics, Creativity, and the Arts*, To appear.

[111] M. L. Wilson, E. H. Chi, S. Reeves, and D. Coyle, "Replichi: The workshop ii," in *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, 2014, pp. 33–36.

[112] N. Zhao, N. W. Kim, L. M. Herman, H. Pfister, R. W. Lau, J. Echevarria, and Z. Bylinskii, "Iconate: Automatic compound icon generation and ideation," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[113] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein, "Hype: A benchmark for human eye perceptual evaluation of generative models," *Advances in neural information processing systems*, vol. 32, 2019.