Towards Better User Studies in Computer Graphics and Vision

Other titles in Foundations and Trends® in Computer Graphics and Vision

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends
Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu and Jianfeng Gao
ISBN: 978-1-63828-132-0

Semantic Image Segmentation: Two Decades of Research
Gabriela Csurka, Riccardo Volpi and Boris Chidlovskii
ISBN: 978-1-63828-076-7

Video Summarization Overview
Mayu Otani, Yale Song and Yang Wang
ISBN: 978-1-63828-078-1

A Comprehensive Review of Modern Object Segmentation Approaches
Yuanbo Wang, Unaiza Ahsan, Hanyan Li and Matthew Hagen
ISBN: 978-1-63828-070-5

Deep Learning for Image/Video Restoration and Super-resolution
A. Murat Tekalp
ISBN: 978-1-68083-972-2

Deep Learning for Multimedia Forensics
Irene Amerini, Aris Anagnostopoulos, Luca Maiano and Lorenzo Ricciardi Celsi
ISBN: 978-1-68083-854-1

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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Z. Bylinski *et al.*. *Towards Better User Studies in Computer Graphics and Vision.* Foundations and Trends® in Computer Graphics and Vision, vol. 15, no. 3, pp. 201–252, 2023.

ISBN: 978-1-63828-173-3

Foundations and Trends® in Computer Graphics and Vision
Volume 15, Issue 3, 2023

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2023, Volume 15, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

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.

Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka and Yile Zhang (2023), "Towards Better User Studies in Computer Graphics and Vision", Foundations and Trends® in Computer Graphics and Vision: Vol. 15, No. 3, pp 201–252. DOI: 10.1561/0600000106.

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.

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: 10.1177/14614448221089604.

[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.