PowerMeter - User Friendly Power Analysis for HCI Studies

Florin Martius
University of Bonn
Bonn, Germany
martius@cs.uni-bonn.de

Lukas Struck
Institute for Computer Science
University of Bonn
Bonn, Germany
struckl@uni-bonn.de

Nele Borgert
Institute of Psychology
University of Bern
Bern, Switzerland
nele.borgert@unibe.ch

Anna-Marie Ortloff
University of Bonn
Bonn, Germany
ortloff@cs.uni-bonn.de

Simon Lenau
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
lenau@cispa.de

Kay Schröder
Hochschule Düsseldorf
Düsseldorf, Germany
kay.schroeder@hs-duesseldorf.de

Theo Sans-Raimbault
University of Bonn
Bonn, Germany
s6thraim@uni-bonn.de

Leonard Heinrich
University of Bonn
Bonn, Germany
s6ldhein@uni-bonn.de

Matthew Smith
University of Bonn
Bonn, Germany
Fraunhofer FKIE
Bonn, Germany
smith@cs.uni-bonn.de

Christian Tiefenau
University of Bonn
Bonn, Germany
tiefenau@cs.uni-bonn.de

Abstract

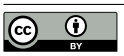
Determining the right sample size for empirical studies is a persistent challenge in human–computer interaction research. Studies with too few participants risk low statistical power and unreliable findings, yet existing tools for power analysis, such as G*Power, are often difficult to use. We present PowerMeter, a user-centered tool that helps researchers estimate appropriate sample sizes for quantitative studies. PowerMeter focuses on usability and interpretability, guiding users through the process of defining key study parameters and understanding the implications of statistical power. In a pilot study ($N = 60$), we find that participants using PowerMeter produced more accurate sample size estimates and reported higher levels of trust and satisfaction than those using G*Power.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

HCI methods; power analysis; statistical program; user study



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799267

ACM Reference Format:
Florin Martius, Lukas Struck, Nele Borgert, Anna-Marie Ortloff, Simon Lenau, Kay Schröder, Theo Sans-Raimbault, Leonard Heinrich, Matthew Smith, and Christian Tiefenau. 2026. PowerMeter - User Friendly Power Analysis for HCI Studies. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3772363.3799267

1 Introduction and Related Work

Statistical power analysis is a foundational step in conducting quantitative studies: before collecting data, researchers should estimate how many participants they need to reliably detect an effect with a desired level of statistical power, given a significance level. Power is defined as the probability of correctly rejecting a null hypothesis when a true effect exists. In other words, it describes the probability that a study will detect a real effect if one is present, which is expressed as $(1 - \beta)$. Studies with insufficient power are more likely to miss real effects. At the same time, underpowered studies are more likely to produce exaggerated and non-reproducible effect sizes, undermining the reliability of statistically significant findings [1]. This variability, together with selective reporting and strict significance thresholds (e.g., $\alpha = .05$), can increase the likelihood that statistically significant findings are false positives and fail to replicate.

Prior work has documented that CHI papers vary widely in sample sizes and often rely on community "local standards" rather than transparent, study-specific justifications [2]. In addition, more recent meta-work highlights that prospective power analysis is rarely

reported in CHI papers that conduct quantitative studies [8, 12], leading to many underpowered studies [12]. Ortloff et al. hypothesize this might be because power analysis requires users to navigate complex interdependent parameters (e.g., effect size, effect size types, study design type, variance assumptions), which can be challenging without statistical training [12].

Since the early power and sample size tables were published, e.g., by Cohen [4], power analysis and the tools supporting it have evolved. Today, researchers can choose from many power analysis tools: web-based calculators[1], programming libraries like the `pwr` package [3] in R or the `statsmodels` library [17] in Python, and stand-alone software such as G*Power [9]. However, all of these require statistical knowledge, in some cases, additional knowledge of a programming language and none offer much guidance.

In addition to these specialized tools, recently, LLMs are increasingly used in HCI research and are becoming part of the CHI landscape [14], with proposed use cases such as: literature review and management [19, 20], qualitative analysis [10], and writing [15]. At the same time, researchers have raised concerns about appropriateness, transparency, and the risk of shallow or ungrounded outputs when LLMs are used in research processes [16]. Outside CHI, recent work has begun to quantify LLM performance on tasks adjacent to statistical planning, including sample size estimation, and emphasizes the need for expert oversight before relying on such outputs [18]. These developments motivate evaluating LLMs not only on perceived usability, but also on empirical reliability in concrete statistical tasks.

In this paper, we introduce **PowerMeter**, a web-based workflow-driven service for power analysis. It supports HCI researchers in determining how many participants are needed before collecting data, providing practical guidance instead of a statistics-first interface. PowerMeter structures power analysis as a step-by-step process, and offers contextual explanations and visualizations that help users understand how design choices affect power. PowerMeter was evaluated against G*Power, which was identified as the sole tool mentioned in a preliminary study involving 12 participants, and the general-purpose LLM ChatGPT 4o. Our results show that PowerMeter leads to higher self-efficacy in conducting power analyses and is preferred by most participants compared to G*Power. When compared to ChatGPT, participant preferences were more evenly split between ChatGPT and PowerMeter. However, our study provides evidence that ChatGPT in version 4o is unreliable for power analysis in our task setting, showing the need for dedicated tools in this field. Together, these findings suggest that workflow structure and just-in-time explanations can improve both performance and users' confidence in their own analyses, while purely conversational assistance may feel accessible but remains risky for supporting rigorous statistical reasoning.

Contributions. This paper makes two contributions. It presents the following:

(1) **PowerMeter**, a workflow-based tool for power analysis that emphasizes interpretability, progressive disclosure, and transparent reporting.
(2) A **controlled user study** comparing PowerMeter with G*Power and ChatGPT, showing improved usability and preference over G*Power, nuanced trust dynamics, and evidence of unreliability for LLM-based statistical assistance in our scenario.

[1] e.g., powerandsamplesize.com or jakewestfall.org/power

2 PowerMeter Prototype

Existing power-analysis tools such as G*Power often require researchers to translate their study plans into a dense configuration of parameters and to specify expected effects as standardized effect sizes (e.g., Cohen's *d*) [5]. In applied HCI research, however, researchers typically reason about effects in raw measurements (e.g., milliseconds, error rates, response scale differences, or success-rate differences). This mismatch makes it difficult to choose plausible inputs and to understand the implications of underlying assumptions, particularly for researchers with limited experience in statistics and power analysis.

PowerMeter addresses these challenges through a guided workflow that treats power analysis as a sequence of small, explainable decisions rather than a single configuration-heavy task. The workflow is structured as a step-by-step process with a progress indicator, helping users maintain orientation. PowerMeter avoids specifying defaults; instead, it offers guidance for users to specify all inputs explicitly. This makes all relevant statistical assumptions visible and encourages deliberate choices, including the controversial alpha significance level of 0.05 [7, 13, 21].

To support informed decision-making, PowerMeter provides just-in-time explanations placed next to the relevant parameters. These explanations focus on practical interpretation instead of formal statistical definitions. Crucially, the system allows users to specify effect sizes not only in standardized form, but also as unstandardized, domain-specific quantities such as raw mean differences or success-rate differences. This enables researchers to reason in terms that align with their measures and study goals.

PowerMeter further supports sensitivity analysis by visualizing how statistical power and required sample size change with varying effect size. This makes the consequences of optimistic or conservative assumptions explicit and helps users assess the robustness of their study plans. At the end of the workflow, the system presents a transparent summary of all assumptions and results, including a power curve and exportable LaTeX text suitable for meta-analysis and replication as suggested by Ortloff et al. [11]. This reporting support reduces friction when documenting power analyses in papers and preregistrations and encourages communication of methodological choices. Screenshots of the PowerMeter prototype can be found in the Appendix, Figures 3, 4, 5, and a screenshot of G*Power in Figure 6. A video showcasing the walkthrough of the power analysis with PowerMeter can be found in the supplementary material.

We informally pretested an early prototype at an HCI conference with $N = 12$ researchers. Participants particularly valued the step-by-step workflow, the visualization of how assumptions affect required sample size, and the automatically generated report snippet. At the same time, several more statistically experienced researchers expressed a desire to specify standardized effect sizes directly; in response, we added support for both inputs in the final system.

3 User Study

With the goal of facilitating power analysis, we tested our refined prototype against two alternatives: G*Power, which was the only power analysis tool named by researchers in the informal study, and ChatGPT (GPT 4o). We included ChatGPT because it is widely used in academia and can readily generate power analyses, making it a plausible alternative that researchers might consult.

We recruited 60 German computer science students enrolled in a lecture on human-computer interaction (HCI) research methods and statistics, which included the basic concepts of power analysis. Although they had received formal instruction in these methods, they had little practical experience applying power analysis. As such, they serve as a proxy for early-career researchers who are beginning to use power analysis in their own work and therefore reflect our intended target group. Participants were compensated with bonus points towards their final exam, and the study protocol was approved by the institutional ethics board.

Participants completed a power-analysis task based on a realistic HCI study involving System Usability Scale (SUS) scores. They were asked to estimate an expected effect size and determine the required sample size for a two-sided mean-comparison t-test. Desired power and α were provided; the full study description is included in the Appendix A.1. Participants had a fixed time budget of 30 minutes per tool and completed a post-study survey assessing self-efficacy, confidence, and trust on a 10-point Likert scale. The Computer Self-Efficacy (CSE) scale is based on Compeau et al. [6]. The questionnaire is provided in Appendix A.2. Since we adjusted items to fit the specific context of our study, we validated the modified 24 items ($N_{items} = 24$) in an independent sample ($N_{prestudy2} = 84$) to ensure psychometric quality.

We used a mixed design, in which PowerMeter was used by all participants and the second tool (G*Power or ChatGPT) varied between groups, enabling within-subjects comparisons between PowerMeter and the other tools, but no direct comparison of G*Power with ChatGPT, which was not our focus. This design allowed us to compare PowerMeter while controlling for individual differences in statistical experience. We further randomized the order between PowerMeter and the alternative tool.

4 Results

4.1 Self-Efficacy and Preference

Figure 1 (left) summarizes participants' (CSE) scores by tool, indicating descriptively higher self-efficacy for PowerMeter and ChatGPT relative to G*Power. Beyond self-efficacy, we observed distinct trust dynamics (Figure 2). Participants reported comparatively *low trust in configuring inputs* in G*Power but *high trust in its output*, suggesting that G*Power's reputation as a "correct" tool may lead users to accept results even when they do not feel confident about the setup. In contrast, PowerMeter elicited higher trust in the input process and comparable or higher trust in the results, compared with G*Power. ChatGPT received similar input trust to PowerMeter, with result trust was slightly lower, but still similar to G*Power

Tool preference patterns further illuminate these trade-offs (Figure 1 (left)). Most participants preferred PowerMeter over G*Power (84%), aligning with PowerMeter's self-efficacy advantage. When the alternative was ChatGPT, preferences were more evenly split between PowerMeter and ChatGPT (58% preferring PowerMeter). However, our evaluation of the generated outputs indicated that ChatGPT was *unreliable for statistical use* in this context, reinforcing the need to distinguish perceived ease-of-use from dependable correctness in methodological tooling.

4.2 Reliability and Accuracy

Regarding the trustworthiness of ChatGPT, although participants generally reported positive experiences, our results show that its outputs should not be relied on without careful verification. While ChatGPT often selected seemingly appropriate parameters, such as the statistical test, test direction, alpha level, and power, when analyzing the participants' chats, we found that ChatGPT's reasoning for the chosen effect size was often problematic. In many cases, the explanation provided did not align with the selected effect size. Specifically, ChatGPT often defaulted to a Cohen's d of 0.5 when suggesting an effect size and calculated the required sample size based on that assumption. When prompted to justify this choice, it typically referred to mean differences and standard deviations commonly associated with SUS scores, but then assumed arbitrary mean differences to retroactively fit a Cohen's d of 0.5. In other instances, ChatGPT produced mathematically incorrect sample size calculations.

More substantial issues emerged in participants' selection of test direction, alpha level, power, and statistical test when using PowerMeter and G*Power. Table 1 summarizes these results. We did not evaluate whether the chosen effect size itself was appropriate, as effect size selection is ultimately at the discretion of the researcher. However, in the scenario we provided to the participant, choices regarding alpha, power, test direction, and test type can be objectively correct or incorrect, given that the instructions of the scenario precisely addressed these. For G*Power, participants selected the correct statistical test in 69% of cases (20/29) and the correct test direction in 72% (21/29). Alpha was specified correctly in all cases, and power in 83% (24/29). However, only 48% of participants (14/29) configured all parameters correctly; in 52% of cases, at least one parameter affecting the power calculation was set incorrectly.

Participants using PowerMeter performed slightly better overall. They selected the correct test in 78% of cases (47/60) and the correct test direction in 88% (55/60). Alpha was correct in all cases, and power in 98% (59/60). All parameters were simultaneously correct in 73% of cases (44/60).

For ChatGPT, parameter selection was largely correct. In two cases, it suggested an inappropriate statistical test (a Wilcoxon rank-sum test and a matched-pairs t-test). In all other cases, parameters matched the scenario, resulting in 97% of configurations being fully correct. Overall, these findings show that even when the study design is clearly specified, configuration errors occur across tools. It is noticeable that especially the test direction (one-sided vs two-sided, also called *tailed*) was a source of errors. While there are many options in G*Power, in its current version, PowerMeter offers only "Means and Median" or "Counts and Proportions". Although it was explicitly stated that a two-tailed, mean-comparing test should be used, many participants failed to do so.

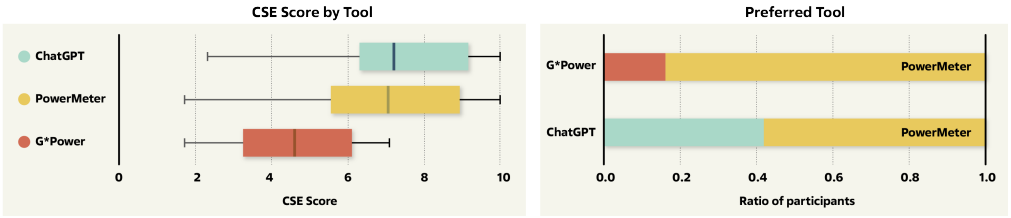


Figure 1: Left: Results of the adjusted Computer Self-Efficacy Scale by tool. 0 means lowest self-efficacy, 10 means highest self-efficacy. Right: Tool preference as chosen by participants.

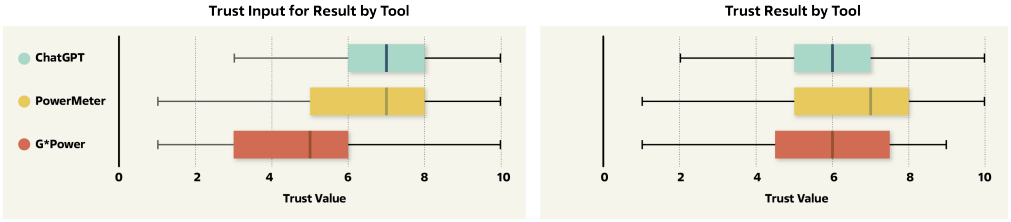


Figure 2: Trust towards the tool used, 0 means lowest trust, 10 means highest trust. Left: Trust that the user inputs were correct. Right: Trust that the tool gave a correct result.

Table 1: Comparison of G*Power and PowerMeter Performance (Percent Correct)

	G*Power	ChatGPT	PowerMeter
Total	29	31	60
Test Correct	20 (69%)	29 (94%)	47 (78%)
Tails Correct	21 (72%)	31 (100%)	53 (88%)
Alpha Correct	29 (100%)	31 (100%)	60 (100%)
Power Correct	24 (83%)	31 (100%)	59 (98%)
All Correct	14 (48%)	30 (97%)	44 (73%)

5 Discussion

Our findings suggest that structured guidance can help users navigate the complexity of power analysis. In particular, PowerMeter's workflow reduces configuration uncertainty and supports understanding of effect sizes and assumptions. The trust patterns observed with G*Power (low trust in input configuration but high trust in output) point to a potential risk: users may rely on authoritative tools without developing an understanding of whether the configuration matches their study design. These results suggest that explanations provided directly within the tool are associated with better understanding of input parameters and, consequently, greater trust in the correctness of those inputs. Given that G*Power does not offer in-tool explanations and instead relies on a separate, highly technical manual, it would likely benefit from integrating contextual explanations, particularly for users without strong statistical backgrounds.

Meanwhile, although chatbots may feel easier to use, our results caution against treating LLM outputs as reliable statistical advice without safeguards, transparency mechanisms, or expert verification. Although ChatGPT produced reasonable values in some cases but arbitrary values in others, LLMs may be better suited to acting as guided assistants rather than autonomous decision-makers in power analysis. Specifically, they could support users by explaining concepts, highlighting test assumptions and guiding them through the process rather than determining critical parameters such as effect sizes. Given their usability advantages, future work could explore integrating an LLM into our tool in a supportive, constrained role to enhance user understanding without compromising methodological rigor. In our specific scenario, ChatGPT was suitable for selecting parameters that matched the study design. However, it remains to be tested whether this level of appropriateness generalizes to other statistical tests and study designs.

While participants using PowerMeter showed a lower error rate in selecting the correct test than those using G*Power, the 78% success rate in selecting the correct test still indicates substantial room for improvement, particularly given that users had to choose between only two tests. Consequently, we extended our tool with explicit support for test selection in the form of an interactive decision tree. Notably, participants had no incentive beyond completing the study to provide correct answers, which may have further biased the results toward lower accuracy.

Our study has design-inherent limitations, such as the use of a student sample and a constrained task scenario rather than participants conducting power analysis for their own real projects. Future work should evaluate PowerMeter in longitudinal deployments, including study planning for ongoing research, and investigate how workflow guidance affects learning and reporting practices over time.

6 Conclusion

Sound power analysis is essential to the credibility of quantitative HCI research, as it determines whether studies can detect meaningful effects and yield interpretable, reproducible results. Despite its

importance, power analysis remains a challenging step in quantitative HCI research. PowerMeter demonstrates that workflow-based guidance, combined with interpretable inputs of effect size, can improve researchers' self-efficacy and user preference compared to established tools such as G*Power, while offering more reliable support than purely conversational assistance. We see PowerMeter as a step toward human-centered statistical tooling that strengthens transparency and methodological rigor in HCI.

Acknowledgments

We thank our study participants and Antonia Sistig for their valuable support in this project.

References

[1] Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience* 14, 5 (May 2013), 365–376. doi:10.1038/nrn3475

[2] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. doi:10.1145/2858036.2858498

[3] Stephane Champely. 2020. Pwr: Basic Functions for Power Analysis. doi:10.32614/CRAN.package.pwr

[4] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). L. Erlbaum Associates, Hillsdale, N.J.

[5] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. doi:10.1145/1219092.1219093

[6] Deborah R Compeau and Christopher A Higgins. 1995. Computer self-efficacy: Development of a measure and initial test. *Manag. Inf. Syst. Q.* 19, 2 (June 1995), 189–211.

[7] Pierre Dragicevic. 2015. *HCI Statistics without p-values.* Research Report RR-8738. Inria. 32 pages. https://inria.hal.science/hal-01162238

[8] Alexander Eiselmayer. 2020. Supporting the Design and Analysis of HCI Experiments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu, USA, 8 pages. doi:10.1145/3334480.3375038

[9] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. doi:10.3758/BF03193146

[10] Elisabeth Kirsten, Annalina Buckmann, Leona Lassak, Nele Borgert, Abraham Mhaidli, and Steffen Becker. 2025. From Assistance to Autonomy – A Researcher Study on the Potential of AI Support for Qualitative Data Analysis. arXiv:2501.19275 [cs.CY] https://arxiv.org/abs/2501.19275

[11] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 483, 28 pages. doi:10.1145/3706598.3713671

[12] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. 2023. SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023).* USENIX Association, Anaheim, CA, 341–359. https://www.usenix.org/conference/soups2023/presentation/ortloff

[13] Yuko Y Palesch. 2014. Some common misperceptions about P values. *Stroke* 45, 12 (2014), e244–e246.

[14] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726

[15] Raquel Breejon Robinson, Alberto Alvarez, and Elisa D. Mekler. 2024. How to Write a CHI Paper (Asking for a Friend). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24).* ACM, New York, NY, USA, 8. doi:10.1145/3613905.3644051

[16] Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, and Sarita Schoenebeck. 2025. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 481, 17 pages. doi:10.1145/3706598.3713120

[17] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. *SciPy 2010* (2010). doi:10.25080/Majora-92bf1922-011

[18] P. Sebo et al. 2025. ChatGPT's performance in sample size estimation. *PLOS Digital Health* (2025).

[19] Tom Völker, Jan Pfister, Tobias Koopmann, and Andreas Hotho. 2024. From Chat to Publication Management: Organizing Your Related Work Using BibSonomy & LLMs. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval.* ACM, Sheffield United Kingdom, 386–390. doi:10.1145/3627508.3638298

[20] Jiyao Wang, Haolong Hu, Zuyuan Wang, Song Yan, Youyu Sheng, and Dengbo He. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–18. doi:10.1145/3613904.3641917

[21] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. 2019. Moving to a world beyond "p< 0.05". 19 pages.

A Appendix

A.1 Study Material

Study Planning Task: Sample Size Required to Compare SUS Scores
Imagine you are planning a study in which you compare two versions of a file manager:

- Version A: Graphical User Interface (GUI)
- Version B: Command-Line Interface (CLI)

Each participant is randomly assigned to one of the two versions. You want to determine whether usability, measured using the System Usability Scale (SUS), differs significantly between the two versions. Your goal is to determine the total number of participants required to achieve a statistical power of at least 80% with a significance level of $\alpha = 0.05$ when comparing means. Please use a two-sided test.

Your tasks:

(1) **Use the software mentioned in the survey to determine the required sample size for this study.** Once you have determined a sample size with which you would conduct the study, please proceed with the survey.

A.2 Study

Demographics

(5) Please indicate your gender
   - Male
   - Female
   - Non-binary
   - Prefer not to say

(6) How old are you? *Text input*

(7) Please indicate your occupation: *Text input*

(8) How many power analyses have you conducted before this study?
   - 0
   - 1-5
   - 6-10
   - 11+

(9) Please briefly describe in what context you have previously conducted power analyses: *Text input*

(10) Would you like to add anything else? *Text input*

Tool Evaluation

This block is repeated twice, once with each tool. One of the tools is

*always PowerMeter, the other is either ChatGPT, or G*Power. The order is random.*

Please conduct a power analysis for the printed scenario in front of you. Use the program "{Tool}" for this purpose. You can find the program on the desktop in the "Tools" folder. The corresponding manual can also be found on the desktop. Additionally, you may use internet research (e.g., Google). You have up to 30 minutes for this task. Once you have determined a sample size with which you would conduct the study, please enter it below and continue with the survey.

(11) How many participants do you need to recruit in total for the study? *Text input*

On the following pages, you will find various situations in which you are asked to conduct a power analysis using {Tool}. Please answer each question in two steps: Do you believe you could perform the task under this condition (Yes / No)? If Yes: How confident are you that you could accomplish this? (1 = "Not at all confident" to 10 = "Very confident").

(12) I could conduct a power analysis with {Tool}... ...if no one told me what to do next.
- Yes
- No

(13) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(14) I could conduct a power analysis with {Tool}... ...if I had never used such software before.
- Yes
- No

(15) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(16) I could conduct a power analysis with {Tool}... ...if only the software manual were available to me.
- Yes
- No

(17) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(18) I could conduct a power analysis with {Tool}... ...if I had seen someone else use it before trying it myself.
- Yes
- No

(19) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(20) I could conduct a power analysis with {Tool}... ...if I could ask someone for help if I got stuck.
- Yes
- No

(21) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(22) I could conduct a power analysis with {Tool}... ...if someone had helped me get started.
- Yes
- No

(23) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(24) I could conduct a power analysis with {Tool}... ...if I had plenty of time to complete the task for which the software is designed.
- Yes
- No

(25) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(26) I could conduct a power analysis with {Tool}... ...if only the built-in help features were available to me.
- Yes
- No

(27) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(28) I could conduct a power analysis with {Tool}... ...if someone had first shown me how to do it.
- Yes
- No

(29) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(30) I could conduct a power analysis with {Tool}... ...if I had previously used similar software to accomplish the same task.
- Yes

- No

(31) How confident are you?
- 1 - Not at all confident
- 2 ... 4
- 5 - Moderately confident
- 6 ... 9
- 10 - Very confident

(32) How confident are you that the inputs you made in {Tool} lead to a meaningful result?
- 1 - Not at all
- 2 ... 4
- 5 - Moderately
- 6 ... 9
- 10 - Very

(33) How much do you trust the result?
- 1 - Not at all
- 2 ... 4
- 5 - Moderately
- 6 ... 9
- 10 - Very

(34) Have you used {Tool} before?
- Yes
- No

(35) Have you used {Tool} for a power analysis before this study?
- Yes
- No

(36) Have you ever received an introduction to {Tool}?
- Yes
- No

(37) Please describe how you chose the effect size used for the power analysis: *Text input*

Tool Comparison

(38) Which tool best supported you in choosing the parameters for the power analysis? *Slider {Tool1} 0 – 10 {Tool2}*

(39) Which tool do you trust the result of more? *Slider {Tool1} 0 – 10 {Tool2}*

(40) Which tool do you find more user-friendly? *Slider {Tool1} 0 – 10 {Tool2}*

(41) Which tool do you find more intuitive? *Slider {Tool1} 0 – 10 {Tool2}*

(42) Which tool do you find more cluttered? *Slider {Tool1} 0 – 10 {Tool2}*

(43) Which tool would you prefer if you had to conduct a power analysis?
- {Tool1}
- {Tool2}

(44) Please describe why. *Text input*

A.3 PowerMeter prototype

Screenshots are on the next page.

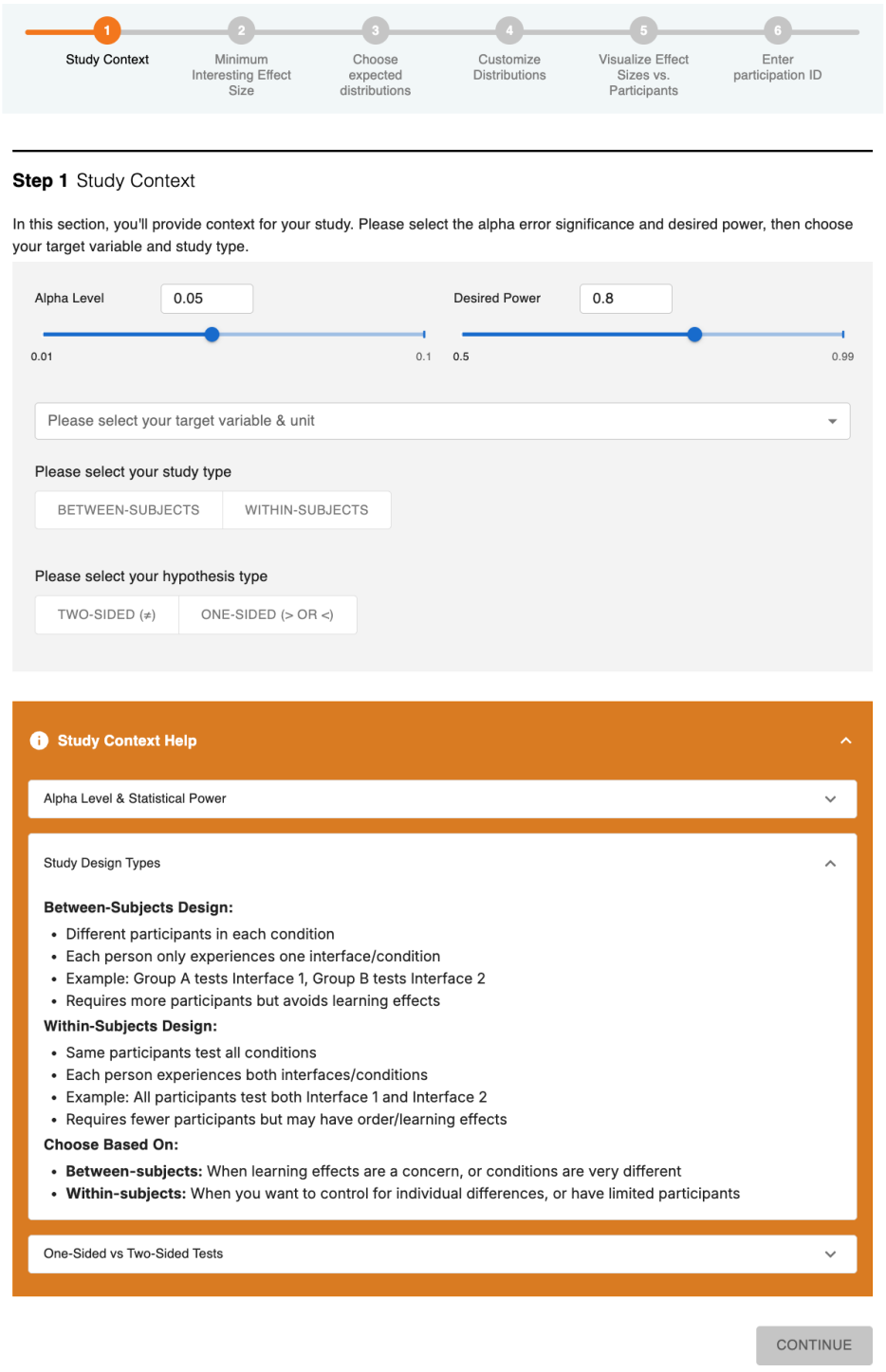


Figure 3: This screenshot presents the study context page of PowerMeter. The form at the top of the page asks for the study context parameters such as the alpha and power values, as well as the study type (between/within) and the hypothesis type (two-sided/one-sided). The target variable and unit are also requested from the user in order to provide additional contextual assistance and to facilitate the determination of the target effect size (see section refsec:powermeter-prototype). The information box located at the bottom of the page offers assistance and contextual information regarding the various sections of the form.

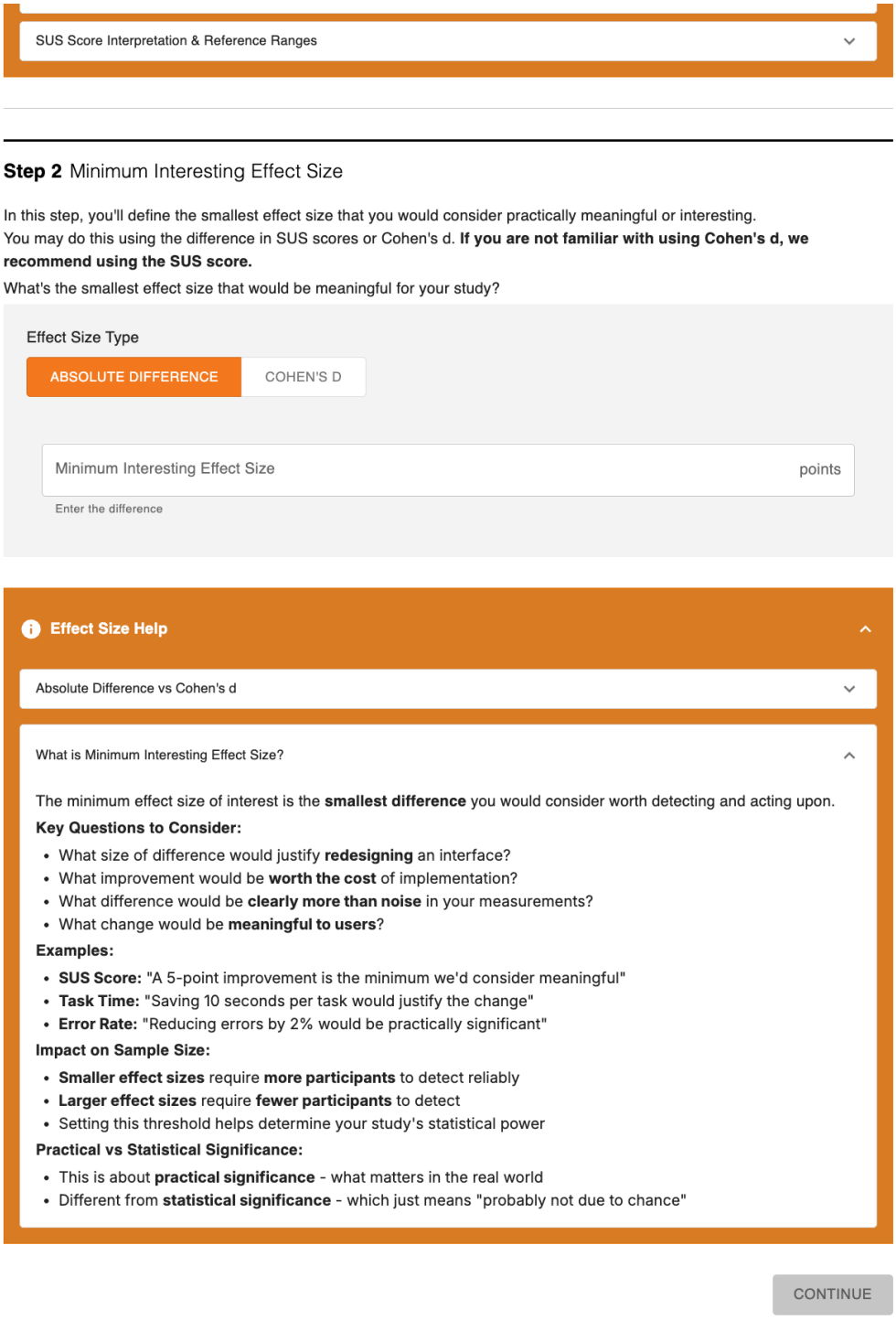


Figure 4: This screenshot presents the minimum interesting effect size page of PowerMeter. The user has the choice between giving the absolute difference in the unit specified in the step before, or Cohen's d [5]. Again, the information box located at the bottom of the page offers assistance and contextual information regarding the different decisions the user has to make in this step.

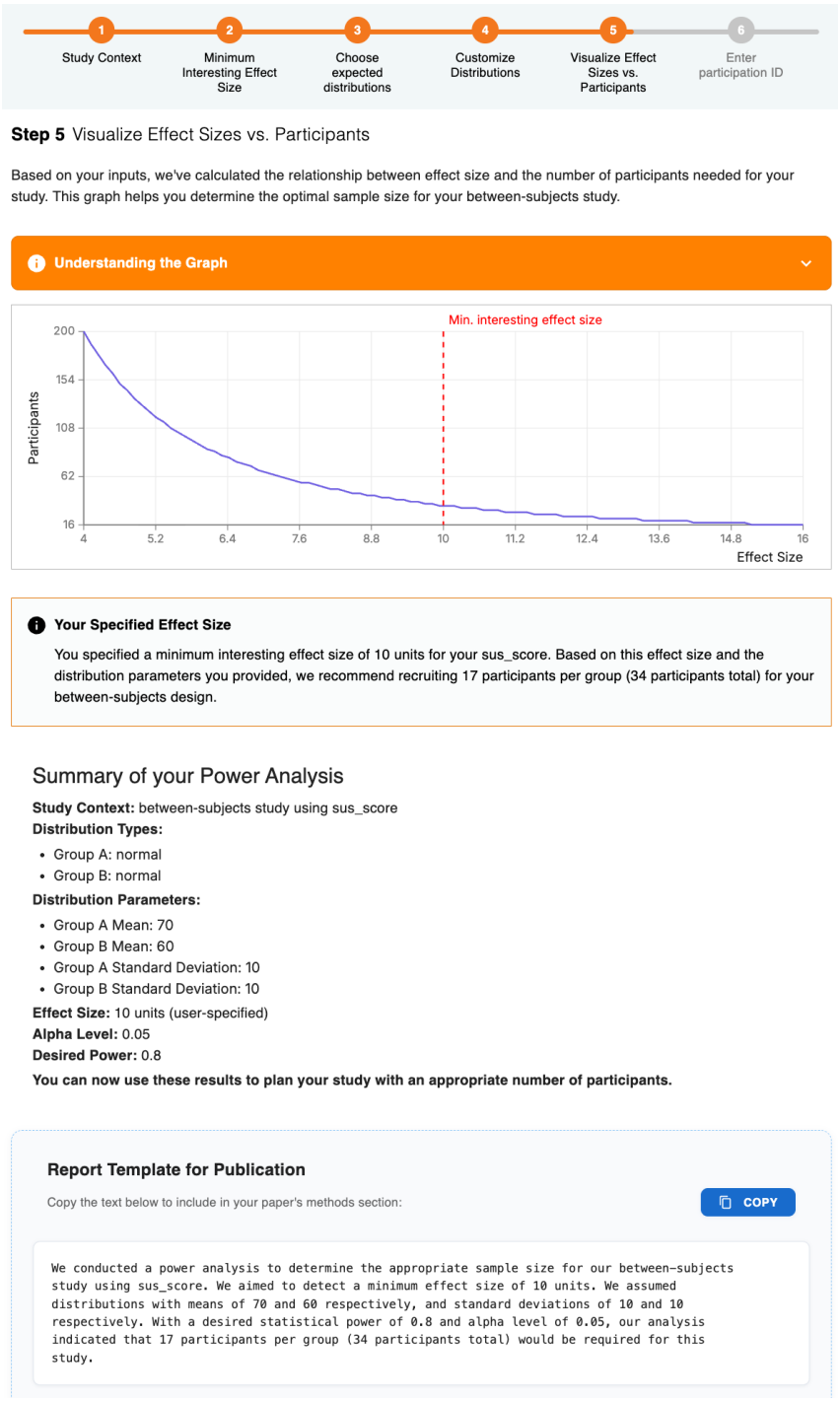


Figure 5: This screenshot presents the results page of PowerMeter. The graph at the top of the page illustrates the relationship between the number of participants required and the minimum measurable effect size, given the selected parameters. The number of participants required can be either read from the graph or from the summary directly below. The summary is divided into two sections. The information box located beneath the graph presents the results in a textual format, while a comprehensive summary of all selected parameters and their respective outcomes is situated below that. As a concluding element, a text segment is provided that may be incorporated directly into a scientific publication without further modification.

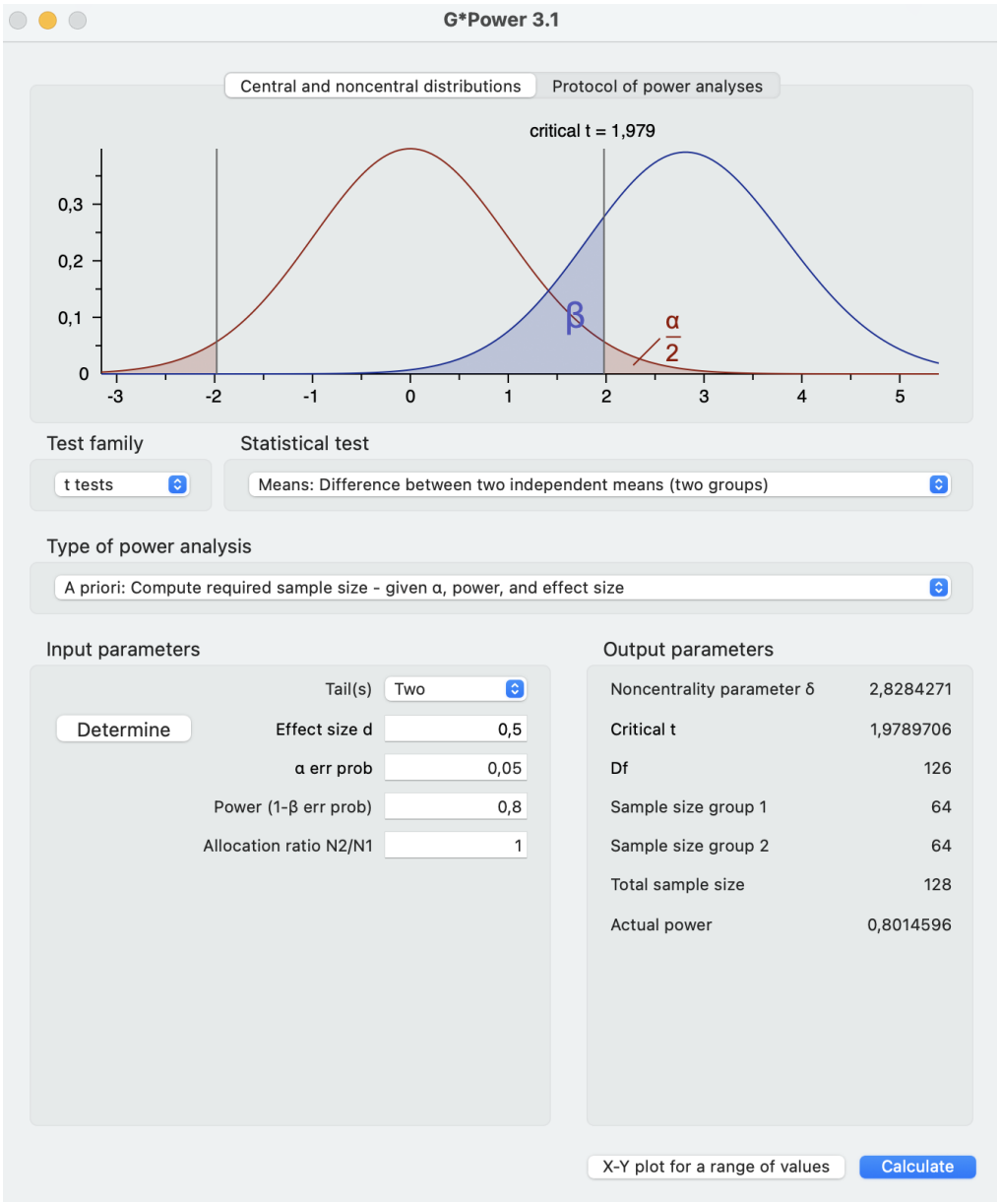


Figure 6: This screenshot presents the G*Power workflow. All parameters are asked in one form, and no contextual information and assistance is given.