# Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation

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

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

Mischa Meier
mischa@mmisc.de
Fraunhofer FKIE
Bonn, Germany

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

Lisa Geierhaas
geierhaa@cs.uni-bonn.de
University of Bonn
Bonn, Germany

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

## Abstract

Statistical reporting, especially of effect sizes, is at the root of many methodological issues in quantitative research at CHI. Effect sizes are necessary for assessing practical relevance of results, a-priori power analysis, and meta-analyses, but currently, they are often not reported. Interpretations in the context of the study and the research field are also rare. To aid to researchers in reporting and contextualizing their effect sizes within their research field as well as choosing effect sizes for power analysis, we conducted a meta-study of quantitative CHI papers. We extracted statistics from all quantitative CHI papers published between 2019-2023 (N=1692). Based on effect sizes and the papers' CCS categories, we present effect size distributions in 12 CHI research fields. Through an additional qualitative analysis of 67 quantitative CHI'23 publications, we identify five categories of approaches that researchers take when interpreting effect size: Comparing test-specific values, assigning size labels, using a statistical or methodological reference frame, comparing different observations and interpreting for the big picture.

## CCS Concepts

• **General and reference** → Metrics; **Surveys and overviews**; • **Human-centered computing** → *HCI theory, concepts and models*; *Empirical studies in HCI*.

## Keywords

meta-science, effect size, statistical power, reporting, data extraction, statistics interpretation, LLM

**ACM Reference Format:**
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 *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3706598.3713671


*CHI '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713671

## 1 Introduction

Researchers frequently emphasize statistical significance when presenting the outcomes of statistical tests. However, effect sizes are equally if not more important [45], but are reported and discussed far less frequently. This is problematic for two reasons: firstly, knowledge about effect sizes in related work is vital to understanding the practical significance of statistical results and to putting them in context to what is common in the research field. Secondly, knowledge about common effect sizes offers valuable guidance for power calculations.

A priori power analysis is ideally used during the design phase of a study to estimate the sample size necessary to detect a desired effect. Eiselmayer et al. found that only 5 of 519 experimental papers at the ACM Conference on Human Factors in Computing Systems (CHI) used power analysis [44]. In the sub-field of developer-centred usable security, Ortloff et al. [96] found that only 5 of 54 papers used power analysis. As a consequence, of the 20 quantitative studies in their analysis that reported enough information to conduct an analysis, all were underpowered to detect small effects according to Cohen's effect size guidelines [31] and 11 did not even have enough power to detect large effects according to these guidelines. Underpowered studies present a threat to scientific validity since they increase the likelihood of both Type I and Type II errors, overestimate effect sizes, hamper future research and waste research resources [45, 57, 58, 96]. Ideally all experimental studies should conduct an a-priori power analysis [45, 96]. To do this researchers have to decide what the minimum effect size which they want to be able to detect is. Ideally, this is guided by effect sizes from previous similar work. However, since work published at CHI focuses on novel contributions [70, 97] there may not be enough prior work that is closely related. In this case, the power calculation can be guided by meta-analyses of effect sizes in the research subdomain. However, for CHI, this does not exist yet.

Similarly, these meta-analyses would be helpful for the interpretation of effect sizes in the context of the research area. Currently, results are often interpreted using a binary scale - there is an effect or there is no effect. This is usually done by using the 0.05

p-value cut-off criterion [13]. Further interpretation of effect size is mostly limited to using the labels of "small" or "large" from generic guidelines, if it is done at all.

However, generic, domain-agnostic guidelines for judging the size of effects like Cohen's [31] are criticized for not taking into account domain-specific and study-specific information [31, 49, 106, 127]. To remedy this, alternative guidelines have been proposed for various domains [15, 65, 106], including some in HCI [91]. However, for a wide range of subdomains in HCI, no effect size guidelines are available. This is unfortunate since effect sizes can vary quite widely within a discipline, as has been shown for psychology [82]. There is also a lack of guidance on how authors can structure their effect size discussion. The CHI submission guidelines call for transparent reporting to enable replication, which would include effect sizes, but they do not make concrete recommendations or offer examples [24].

To tackle these problems our paper makes several contributions:

1) We created a meta-analysis support tool using an LLM (GPT-4o [94]) to extract and categorize statistics from HTML CHI papers. We focused on those values that are helpful for interpretation, power analysis, and meta-analysis: effect sizes, participant numbers, confidence intervals, and p-values. We built in plausibility checks to ensure the LLM was not hallucinating. We provide our extraction and analysis code[1], so other researchers can replicate our results, or extend it to a subset of research that is relevant to them.
2) Using our tool, we extracted statistics from all CHI publications between 2019 and 2023 that contained statistical tests (1692) and manually checked for inconsistencies. For our meta-analysis, we classified publications into 12 different CHI subdomains and created overviews for standardized effect sizes for each of these research areas. We focus on standardized effect sizes which can be converted into each-other: Cohen's d, Pearson's r, $R^2$, odds ratio, the Common Language Effect Size (CLES), the non-parametric correlation coefficients Kendall's $\tau$ and Spearman's $\rho$, Cramer's V, and $\eta^2$, including variants like $\eta_p^2$ or $\omega^2$. This data will aid researchers in interpreting their effects in the context of their domain and in planning power calculations.
3) To offer further support we analyzed effect size discussions in papers from CHI'23 and identified five categories of approaches taken by researchers. We discuss these and offer a framework to help authors choose which approaches are appropriate for their statistical analysis.

## 2 Related Work

We explain key terms related to inferential statistics, focusing on the two concepts of statistical power and effect size, before summarizing related work on statistical reporting at CHI and other communities and situating our work within the related work in the domain of meta science. Finally, we discuss work related to LLMs.

### 2.1 Theoretical Background on Statistics

Roughly half of CHI publications use quantitative analysis to draw and support conclusions [21, 22] and most use the Null hypothesis significance testing (NHST) framework to do this [66]. Various calls to action have advocated for a switch to using Bayesian methods for inference [66, 70], but as of now, most work still uses NHST [13]. Given that this work seeks to analyze the status quo of effect size (use) in the CHI community, we focus on the methodological framework used most, and will give a brief overview over key terms in NHST, which are relevant to this work, while deferring to Kaptein & Robertson [66] and Kay et al. [70] for criticism of these practices.

NHST is a problematic merger of two separate approaches to quantitative data analysis, one by Fisher and one by Neyman and Pearson [103]. It compares the null hypothesis which often assumes no relationship between independent and dependent variables and the alternative hypothesis, which does [45]. A p-value constitutes the probability of getting results which are at least as extreme as those which were observed in the sample, assuming that the null hypothesis is correct [45]. To control for false positives (type I errors), researchers only accept the alternative hypothesis, when the p-value is below a set significance criterion $\alpha$. Commonly, $\alpha$=0.05 (5%) is used, however, this threshold and p-values in general are also a critically debated topic [117, 150]. The probability of a statistical test correctly rejecting the null hypothesis is called statistical power [31]. Commonly, researchers strive to achieve a power of 0.8 [31, 38], which means that an actual effect will be detected 80% of the time, while a type II error (false negative) occurs 20% of the time.

The decision to reject the null hypothesis is binary, but research results lie along a continuum. The effect size, which measures the strength of association between independent variables and dependent variables [45] represents this. On one hand, effect sizes can be presented in the units of the dependent variable [6] or as differences in percentages or amounts, which we refer to as simple effect sizes. On the other hand, unitless effect sizes abstract from the units of the study, e.g. by using normalization with the sample variance [6]. Within unitless effect sizes, those related to group differences (the d-family) are distinguished from measures of association (the r-family) [116] and risk estimates [49]. Finally, since the effect size reported with a study is a point estimate of a population characteristic, confidence intervals (CIs) communicate the uncertainty around these estimates. CIs become more narrow, the less uncertainty there is. The 95% CI is the most frequently used, and it includes the true population value of the estimated parameter in 95% of the samples [45].

The most common type of power analysis is a-priori power analysis, where the necessary sample size is estimated from $\alpha$, power, and the effect size. This is an important step in planning a scientific study, as underpowered studies may waste resources, since they are likely to fail to reject the null hypothesis [45], and non significant results are harder to publish [5, 122, 123]. Overpowered studies may be overly sensitive and able to detect very small effects that are not relevant in practice [45]. Additionally, when focusing only on p-values to interpret the results of statistical tests, both underpowered and overpowered studies can lead to misinterpretations of the practical relevance [45].

Most power analysis software, like GPower [48] uses unitless effect sizes in the procedures for a-priori power analysis, although depending on the type of test, it may enable input of descriptive statistics which are then used to calculate the unitless effect size.

[1] https://github.com/Behavioral-Security/A-meta-study-of-effect-sizes-at-CHI

Several strategies can be employed to determine which size of effect to use in a priori power analyses. Estimates can either be based on similar studies from the literature [45], researchers' judgments of the smallest effect size of interest [3, 76]. or guidelines for judging the size of effects [45]. The most frequently used of these are Cohen's guidelines, whereby a Cohen's d standardized effect size of 0.2, 0.5 and 0.8 constitute small, medium and large effects respectively, and for Pearson's r, 0.1 is judged to be a small, 0.3 a medium, and 0.5 a large effect [31]. Guidelines for risk based effect sizes are not as commonly cited, likely because their interpretation is highly dependent on the base rate of risk in the population, i.e. when the base rate of risk is higher, then interventions lowering this risk are more relevant. Other guidelines exist, but are not as frequently referenced, e.g. Ferguson describes minimum thresholds for practically relevant effect sizes of different types in Table 1 [49]. These guidelines have been criticized for not taking into account context and methodological issues [31, 49, 122, 127]. To remedy this, various domain specific guidelines have been proposed, both inside the CHI-community [91], and in adjacent disciplines like psychology [122].

## 2.2 Meta Research

Meta research is the study of research, which encompasses investigating research methods, reporting practice, reproducibility of research, evaluation methods, and how to incentivize good science [59]. Meta research has received much attention in the wake of the replication crisis, whereby many scientific results have been shown to be difficult or impossible to reproduce [34], most famously in psychology [93] and medicine [57], but also in other fields, such as economics [12].

This work contributes to the meta-research areas of methods and reporting. Our methodological contribution is a process to extract statistical values from scientific publications. Our contribution to the domain of reporting is an overview over the range of effect sizes common in various research areas at CHI and an analysis of approaches used when interpreting them.

Other methodological meta research in HCI and CHI has investigated aspects of research methods from various stages of the research process, from literature search [98], recruiting [119, 131], types of participants [81, 88] and study methods [40, 111, 119] to data analysis of quantitative [70, 84, 87, 110, 121, 147], and qualitative data [95]. Other prior work examined the definition of HCI as a discipline through the lens of problem solving [97], and investigated meta-analysis [63, 64, 155] and systematic reviews [114] as well as practices supporting research quality, such as pre-registration [30] and open science [39, 50, 101].

Various aspects of reporting have been explored at CHI and in HCI, including reporting of participant compensation [102], openness and transparency [7, 120] and race and ethnicity [23]. On the meta information level, Bd et al. discuss issues with authorship order and propose an interactive solution [9]. Liu et al. explore decision-making during study design and data analysis, and discuss visualizations for communicating such decisions [84]. Research investigating visualization of uncertainty [55] and presentation of effect size [72] suggests that currently common ways of presenting statistics, such as confidence intervals [55] or means and mean differences [72] may lead to research readers overestimating effects. There have also been efforts to foster transparent reporting of statistical information [42, 68, 69, 145].

## 2.3 Reporting of Statistical Results

Reporting statistical results accessibly, understandably and completely is important to ensure other researchers can build upon prior work, and form their own interpretation of findings. The CHI guide to a successful submission stresses that "[...] statistical analyses should be described with significant detail" and that such papers "should include enough detail for an independent researcher or practitioner to (1) independently evaluate the correctness validity , and reliability of your [...] analyses and (2) reproduce and replicate [...] experimental methods", in its Transparency section [24]. However, it neither provides more detailed guidance on how to achieve this, nor references guidelines that do so. A special interest group on transparent statistics took place at CHI'16 [68] and the group has since organized further discussions at CHI and IEEE VIS [69, 139, 145]. Work is ongoing on a guideline for transparent statistics reporting in HCI, with the section on effect size currently in alpha, and other sections in the drafting stage [138].

Other disciplines have guidelines on reporting statistics which are endorsed by some publication venues, of which we discuss examples. In medicine, extensive guidelines exist for many specialized variants of study designs [46, 47]. One of the most frequently referenced is the CONSORT checklist [85]. As the term checklist implies, this does not give detailed instructions on how to report information, but rather specifies what to report. Evaluations comparing quality reporting of endorsing venues and non-endorsing venues suggest that it has a positive effect on reporting, although problems with reporting remain [105, 141]. Other medical reporting guidelines are more detailed, e.g. even the abbreviated website version for the AHA/ASA journals specifies the order in which to present statistical values [62]. A more comprehensive version gives some examples about commonly used effect sizes in the application domain and consolidates the CONSORT guidelines within their section on randomized controlled trials [1]. The publication manual of the American Psychological Association (APA) gives detailed instructions regarding various aspects of the publication process, including reporting [4]. The main text remains non-specific due to the variety of possible tests, but the sample tables include examples of the information which should be reported for several commonly used tests in psychology, including t-tests, $\chi^2$-tests, ANOVA and regression [4].

Prior work suggests that reporting statistical results in a complete and transparent way is a challenge that applies to many disciplines related to aspects of Human Computer Interaction (HCI), such as psychology [90], social science [130], software engineering [65], usable security [32, 54, 96], as well as in the domain of HCI itself [22, 89].

Salehzadeh Niksirat et al. investigated reporting practices at CHI, including of statistics, and found that reporting of descriptive statistics, clearly stating the test procedure, and reporting of test statistics and p-values was largely sufficient, but reporting of effect sizes and even more so, confidence intervals was lacking [120]. Besançon and Dragicevic focused on the communication of results as

dichotomous, i.e. using significant vs. not significant as a proxy for the results being interesting or relevant and found that phrasings related to dichotomy are common at CHI [13]. Dragicevic et al. propose reporting multiple possible analyses, i.e. multiverse analyses, to promote transparency of statistical reporting [42].

## 2.4 Large Language Models

Language models are statistical models which predict the next word or token in a series of words by assigning a probability to each sequence of words [11]. Large Language Models (LLMs) may use billions of parameters and are trained on huge datasets [17, 37, 137]. This means that they are able to generate natural language convincingly and do not require as much topical fine-tuning of parameters as prior models, and can be tuned using prompts [153]. The most popular architecture is currently the transformer architecture [133, 143].

Research has investigated various use cases for generative AI like LLMs, e.g. in information extraction [26, 27, 80, 107], annotation [135, 156] or reasoning [28, 146], and more domain-specific in education [10, 67, 136], as a programming aid [10, 86, 142], and in science [134, 144, 148].

We focus in particular on work using language models to extract data from scientific publications. Lee et al. developed a BERT-based model pretrained on biomedical text [80]. Circi et al. explored the capabilities of three different LLMs: GPT-3.5 and -4, and Claude to extract sample information about polymer nanocomposites [26]. They used prompts and provided a JSON template to fill for the task and reached an F1 score between up to 0.88 for Claude2 and 0.36 for GPT-3.5 [26]. Polak and Morgan extracted material properties from research literature, using prompt engineering with GPT-4 and GPT-3.5, as well as LLAMA2 for an open source comparison to achieve around 90% precision and over 80% recall, when using GPT-4 [107].

While LLMs have many interesting applications and use cases, they also come with disadvantages. Even if the generated content seems convincing and well-phrased, responses can in fact be misleading and contain wrong information [61]. This is called hallucination [61]. In addition, LLMs and AI in general do not actually know or understand the text they generate or the prompts they are given, but generate text based on probability distributions. These are dependent on the training data, and will repeat biases present in the training data [11]. Training data scraped from the internet is not a balanced representation of the global population, over-representing younger people and people from developed countries [11]. This leads to LLMs generating content that exhibits biases against marginalized populations, e.g. replicating stereotypes [8, 11, 20, 157], and that contributes to echo chambers and polarization of debates and opinions [126]. Finally, model development, training, and deployment incurs environmental costs. Increasingly large numbers of parameters also increase the amount of energy needed during training and development [11, 129], as well as increasing $CO_2$ emissions [129] and water consumption, e.g. from cooling in data centers [158].

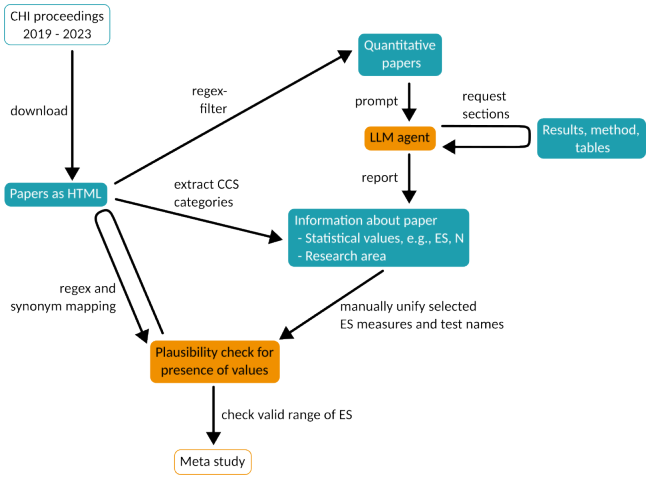


**Figure 1: Overview of Extraction Process**

## 3 Extracting Statistics

In this section we describe our data collection process to extract statistical information from five years of CHI publications. Figure 1 shows an overview of the extraction and analysis process.

We downloaded and filtered the CHI papers, so that only quantitative papers remained (see Section 3.1). For a given paper, we extracted identifying information, i.e. the title and DOI, as well as the abstract and structure of the paper from the HTML to pass to our LLM agent in the next step. As described in Section 3.3, the agent used our prompts and this information to request content from the paper and extract the relevant statistics described in Section 3.2. The final output per paper was converted to JSON and forms the basis of our meta-analysis, which we describe in Section 4. To enable re-use of our data, we provide the scripts used for analysis and our analysis results on Github [2].

### 3.1 Sample of CHI Publications

We selected five years of CHI publications for our analysis (2019 to 2023). CHI was first published in HTML format in 2019. This machine-readable format facilitates extraction considerably. Using optical character recognition (OCR) to generate machine-readable text from PDF is especially error-prone around tables [36], diacritics [14] and when mixing languages [100]. Statistics are often reported in a mix of Greek and Latin letters, with sub- and superscripts and frequently in tables, so using the machine-readable text from HTML directly is more reliable.

Concurring with prior work [21, 22], we define a quantitative paper as one reporting statistical tests. We used regular expressions to identify p-values, CIs, and Bayes factors. P-values and CIs are specific enough to ensure that an actual test was conducted and reported, rather than theoretical papers discussing the merits of statistical testing methods, and at the same time not limited to a single type of hypothesis test in the way that e.g. effect sizes are. We configured our regular expressions to include variations in capitalization and plural and singular terms. For confidence intervals, we

[2] https://github.com/Behavioral-Security/A-meta-study-of-effect-sizes-at-CHI

filtered for singular and plural of use of the term "confidence interval". Since the often used abbreviation "CI" is a letter combination not uncommon in English, we did not use this as a filter. To get an idea of the frequency of use of Bayesian statistics for analysis, we filtered for inclusion of the term "bayes factor", since Bayes factors are a common way to compare models in Bayesian analysis [125]. If a paper satisfied at least one of these criteria, we included it in our sample of quantitative papers. We evaluated our filtering criteria by manually checking a random sample of 145 papers evenly spread over all five years in our sample. In our evaluation, all but two papers were correctly classified, one each as false negative and false positive. False positives may occur due to meta-scientific or theoretical papers discussing statistics, but not using and reporting statistical tests, such as [18]. False negatives can be caused by papers using "p>", "p=" or "p<" for values abbreviated p, which are not statistical p-values, although this did not occur in our filtering ground truth, or by papers reporting statistical tests, but providing statistical values only in images, e.g. [25] reporting the results of a Wilcoxon ranksum test in a Table included as an image.

Between 2019 and 2023, 3724 papers were published at CHI. Of these, we identified 1692 (45%) to be quantitative and used these as the basis for the rest of our work. The ratio of quantitative to qualitative papers is similar to findings in prior work [21, 22]. Of these quantitative papers, 13 (0.8%) were identified as using Bayesian methods, which we did not include in the meta-study.

## 3.2 Statistics of Interest

We focused on extracting the information most relevant when deriving domain-specific guidelines for effect size judgments: The name of the conducted hypothesis test, sample size N, the p-value, effect size and the confidence interval around the effect size. For some types of tests, like regressions with multiple factors or multi-way ANOVAs, multiple p-values, effect sizes and confidence intervals may be associated with a single test.

These statistics are generalizable in the sense that they apply to many tests in the NHST paradigm, which avoids the need for separate handling of every imaginable test. For this reason, we decided to focus on standardized effect sizes. Confidence intervals around the effect size can be used to gauge reliability of the estimates, however, reporting confidence intervals around effect sizes is currently not common at CHI. In their absence, sample sizes can be used to assess reliability. Test names are necessary to get information about the experiment design, as effect sizes may share a name, but be calculated differently for within group compared to between groups designs, as is the case for Cohen's d, which has versions for both. Finally, while p-values are not directly useful for our further analysis, they seem to be the statistic reported most often and serve as an identification criteria of hypothesis tests conducted based on NHST.

## 3.3 Extraction Tooling

We focused our extraction on data available in text form, including the main text, tables and captions but excluding images. Text extraction from images requires OCR which can be inaccurate, especially for non-latin characters [14, 100], and if extracting values represented by visual elements, the accuracy depends on the scale of the image. We did not analyze supplemental material.

We used OpenAI's GPT-4o [94] in combination with a LangGraph-based agent [79] to extract the statistics and used the lowest possible temperature settings to lower randomness in our extraction process. Our prompts are described in more detail in Section 3.4. Three tools were defined for the LLM to utilize: a section read tool, a table read tool, and a test report tool. The read tools allow the agent to request a section or table. The agent can specify the section or table by its index or heading. Sections are returned as plain text, while tables are returned in csv-format, accompanied by a prompt instructing the LLM to read the table row by row. The report tool provides the LLM with a set of predefined fields that must be satisfied with statistical test data, such as the p-value or the effect size measure type, or "UNKNOWN". Upon receiving the initial prompt, the LLM determines which sections to request via the tool. Based on the section text, it identifies the statistical tests and makes use of the report tool. This process is subsequently repeated with the tables.

## 3.4 Prompt Engineering

We built on our experience extracting statistical information from papers, e.g. from creating the ground truth as described in Section 3.5, to derive an initial prompt. We improved on this prompt through manual experimentation, similar to [112, 135, 156]. For example, in the extraction results based on our initial prompt, information from tables was largely missing. In subsequent iterations, we adjusted the prompts and tooling to consider tables separately. Our prompts consisted of a system prompt and a task prompt. The system prompt consisted of a persona description and general instructions including response format. The task prompt had a closer description of the task, including the statistics and information to be extracted and a description of special cases and how to handle missing information. Finally, the prompt included information about the paper which we extracted from the HTML beforehand: the title, abstract and structure, i.e. section indices and headings, as well as table indices and captions. We provide the final prompts in the supplemental material.

## 3.5 Evaluation

We manually annotated a sub-sample of quantitative CHI papers (N=25) to compare our extraction results to. To create this extraction ground truth, we collected a random sample of five quantitative papers per year from our sample. One researcher with a research focus on methodology and statistics manually extracted the relevant statistics and entered them into a JSON data structure so they were human- and machine-readable. To reflect the data available to the LLM, they only used information from the main text, tables, and captions. The researcher corrected typographical mistakes ("p < 05", e.g. missing the decimal point). Context was used where possible to derive values, e.g. sample sizes for an individual test from tables reporting demographics, or from degrees of freedom and the number of conditions in a test [99]. In case of uncertainties the researcher discussed these with their co-authors and noted remaining uncertainties within the ground truth.

For evaluation based on the ground truth, we focused on the effect sizes and that they are reported together with the correct

test and sample size. From the 25 papers, we extracted 528 tests of which 438 did not contain effect sizes. From the remaining 90 tests, we extracted 403 effect sizes. Of these, 130 were relevant for the meta-study. The remaining effect sizes consisted of 258 regression coefficients and 15 correlation coefficients where the type of correlation was not specified. We could not use regression coefficients for the meta-study because additional descriptive statistics, i.e. the standard deviation of the outcome variable, and group sizes, are necessary to convert them to an effect size useful for meta-analysis [83].

Overall, 86 of the 130 effect size values were extracted correctly by the LLM and 2 were extracted with a missing −, i.e. the wrong direction of effect was extracted, which is not critical, since we use absolute values in the meta-study. 42 were missing. Of these, 23 effect sizes were only reported in tables, and not individually addressed in the text, 12 were Cohen's ds reported as part of a meta-analysis, but only in the form "d=", which might not be clearly associated with statistics, and one was reported in the method section, which might not have been requested by the section read tool. The remaining 6 missed effect sizes were clearly identifiable as statistics by us and we do not know why the LLM did not extract them. No effect size values were extracted incorrectly. Having no incorrect effect size values was our main goal, so we accepted the 32% false negative rate

Of the relevant 90 tests we extracted manually, the LLM extracted 64, of which 37 were extracted with correct and complete test names, 9 with correct, but underspecified names (e.g. "t-test" instead of "dependent t-test") and 14 with incorrect names. The test names are only relevant to distinguish between within or between groups designs. For the 64 tests which were extracted by the LLM, participant numbers were extracted correctly in 49 cases and incorrectly in 12 cases, where the LLM did not recognize that a different subsample was used in the test. Three tests were not reported with determinable sample sizes. The sample sizes do not directly affect the effect size, but only the weighting of the effect sizes in the meta-study.

Given that we prefer missing data to wrong data, we considered this sufficient for use in our meta-study and used our extraction procedure on our full sample of CHI papers.

## 4 Constructing Effect Size Guidelines for CHI Research Areas

We converted the effect sizes extracted in the previous section to a common effect size to investigate effect size distributions in 12 research areas at CHI.

### 4.1 Categorization of Papers

To identify papers' research areas for our meta-study, we started with the top level ACM Computing Classification System (CCS) categories. A paper can fall into multiple different categories. *Human-centered computing* was unsurprisingly the most frequent research area by a large margin (3280, or 91.6% of all papers) with the next most frequent category being applied computing (10%). Other categories were also fairly general, making them less useful. To remedy this, we decided to additionally use the subcategories of these general three categories, *general and reference*, *human-centered computing*, and *applied computing*, for more concrete and thus interpretable research areas.

For our sample of quantitative CHI papers, this resulted in 141 different categories. In the following we describe the categorization for a subset of these, the 599 papers, which reported effect sizes which we used in our meta-study. For the most frequent CCS categories, we merged two pairs of categories which were conceptually similar. We merged *user studies* and *empirical studies in hci* as well as *empirical studies in collaborative and social computing* and *collaborative and social computing*. We then chose the top 10 most frequent categories, which were specific enough to allow distinction between papers and covered the largest possible number of these papers. There were two more categories which appeared the same number of times as the least frequent of these categories, in 27 papers, so we included these to avoid excluding categories due to alphabetic ordering, and not different frequency of appearance. This resulted in 12 research areas. While we extracted statistics and conducted our meta-analysis for all quantitative CHI papers, we focus analysis per research area on these 12. Only 12 papers of 599 do not fall into one of these areas, they are nevertheless included in our meta-analysis when we discuss effect sizes at CHI overall.

While we would have liked to use CHI-specific categories such as the subcommittees, data on subcommittee assignment is not available publicly. Of the 12 categories we discuss, *collaborative and social computing*, *interaction design* and *security and privacy* match well with the more CHI specific categorization into subcommittees of *Interaction Beyond the Individual*, *Privacy and Security* and *Computational Interaction* respectively. Some matches between these two categorization schemes are not as clear. Our categories of *interaction design* and *human computer interaction (hci)* also fit to some extent with the CHI subcommittees of *Blending Interaction* and *Interacting with Devices*. Our results for these categories are directly applicable for researchers submitting to CHI. The subcommittees of *Accessibility and Aging* and *Learning, Education and Families* have overlap with the category of *social and professional topics* and the same applies to the subcommittee of *Specific Application Areas* and *applied computing*. Our guidelines for these categories may be better suited than those for CHI overall, but researchers should take their own research specifics into account, even more than for some of the other categories. While our research areas of *virtual* and *mixed/augmented reality* are concrete and identifiable for researchers, they lack an equivalent within the subcommittees.

### 4.2 Converting Effect Sizes

For our meta-study, we focused on standardized effect sizes which can be converted into each-other: Cohen's d, Pearson's r, $R^2$, odds ratio, CLES, the non-parametric correlation coefficients Kendall's $\tau$ and Spearman's $\rho$, and $\eta^2$. In addition, $\eta_p^2$ (partial eta-squared) is equivalent to $\eta^2$ in one-way ANOVA, but not for multi-way designs [75], where all $\eta_p^2$ may add up to values of over 1 [113]. However, judging the size of $\eta_p^2$ is possible through effect size specific benchmarks derived from those set up for Cohen's d [31, 113] and as such, we included it in our calculations. Similarly other variants of $\eta^2$, such as generalized $\eta^2$ ($\eta_G^2$) or less biased variants

like $\omega^2$ may not always yield the same value when applied to the same data set, but they are measured on the same scale as $\eta^2$ and we treated them as such in our calculations. On the other hand, we excluded pseudo-$R^2$ used in logistic regression, such as McFadden's or Nagelkerke's pseudo $R^2$, since they are not measured on the same scale, but their true maximum depends on the data used in the model [115, 128]. Similarly, we excluded $\phi$, since it is only the same as Cramer's V in 2x2 contingency tables, but different in larger tables [33], and based on our extraction, we cannot tell reliably which is the case.

Given that more conversion formulas were available for r, and its value range is clearly bounded in contrast to Cohen's d, where the upper limit is ∞, we converted all effect sizes to Pearson's r. Since we were not interested in the direction of effects, but only the size, we used the absolute value of the reported effect sizes in the conversion. We applied the average correction factors to our conversions, which Poom and af Wåhlberg derive through Monte Carlo simulation to improve the accuracy of converted effect sizes for meta-analyses [108]. For individual cases, this can lead to values outside the valid range for r. We report and incorporate these in our meta-study as is to provide consistency for conversion back to the original effect size. Researchers using the guidelines should refer to Table 1 for the valid range of effect sizes and only interpret them within this range. The conversion formulas we used are also in Table 1.

## 4.3 Data Cleaning

We describe the process of data unification and cleaning in the following.

Some statistical tests have multiple names, a common example being the Wilcoxon rank-sum test, also known as Mann-Whitney U test or Wilcoxon-Mann-Whitney test. In addition, our LLM-extracted data often contained additional information on the variables involved in the test as part of the name. For further analysis, we unified the test names by mapping them to a list of test names we collected from a statistical text book [51]. We added the names of extracted tests which did not correspond with the tests on our list where applicable. We followed a similar procedure for the effect size measures, except that we focused specifically on effect sizes we would use in our meta-study, as described in Section 4.2

In our analysis of the ground truth, we noticed that sometimes, effect sizes were reported twice: once as a factor, and once with the omnibus test, so for our final analysis we removed 6009 rows of data containing duplicates.

Since manual extraction and checking for a ground truth is time and labor intensive, and cannot tell us about performance in the full sample, we also implemented several plausibility checks using regular expressions on the HTML source, focusing on effect sizes and Ns, as we primarily used these values in the meta-study. Due to inconsistencies in how statistical values are formatted, a simple check that matches exact values in the HTML is insufficient. Therefore, we designed regular expressions to capture variations, such as leading zeros, white spaces, and different formula representations, and also created a synonym mapping for statistical measure names, as these are used inconsistently across different authors. These plausibility checks initially flagged 235 effect size values and 110 Ns, which we manually checked. We removed 34 wrong effect size values and 40 wrong Ns from the data set.

As an additional plausibility check on the effect size values, we checked that they were within the limits of a valid effect size of the reported effect size measure, as depicted in Table 1. We removed 43 of 6755 effects for invalid effect sizes. These invalid effect sizes can be due to a hallucination, but can also be introduced into the papers through human error, such as the ones we found during the creation of our human-annotated ground truth, where we saw a correlation coefficient larger than 1.

## 4.4 Data Analysis

We used R [109], specifically the `tidyverse` [151] for data manipulation and `weighted_quantile()` from the `modi` package [56] for the meta-study.

Effect sizes from within- and between groups designs should not be compared directly, since the error variances are calculated differently for different designs, especially in the case of $\eta_p^2$ [92], which was quite frequent in our sample (21.3% of the valid values). Due to our method of extraction, we used information from the extracted test names and in some cases also effect size measures to derive the experiment design, e.g. independent t-tests are conducted for between groups designs and dependent t-tests for within groups designs. However, due to unclear reporting of tests as merely "t-tests" without further specification or various types of regression where determining the use cases in terms of experiment design required disproportional effort, 1379 (20.6%) of effect sizes could not be assigned to an experimental design for certain. Analysis results from within-groups designs have less random error and e.g. the same difference between groups results in a larger effect size. Wrongly counting an actually within-groups effect size from an unclear test as a between-groups effect size means it will likely be comparatively larger than the between-groups effects. Within-groups studies generally have less random error than between-groups studies. Thus, given an effect size reported with an unknown study design, classifying this as between-groups would represent a larger effect (e.g., mean difference) than classifying E as within-groups. Therefore, misclassification of a within-groups effect size as between-groups will shift the overall distribution of the effect sizes to be larger than they actually are. Overestimating effect sizes is already a problem due to publication bias [58]. We believe underestimating effect sizes to be less grave and wanted to retain the information value of the unclear effect sizes and so we counted them as within-groups.

For each of these analyses, we first calculated median effect sizes per paper, weighted using the sample size of each test. We chose to use the median as a measure of central tendency, since we are interested in typical effect sizes at CHI and robustness against outliers. Where sample size was not available, we calculated the median sample size for the paper as a stand-in, and if no sample sizes were extracted for a given paper, we substituted the overall median sample size, 40. For extreme values of sample size > 352.5, using Tukey's fences as the outlier criterion [140], we substituted the maximum non-outlier value: 352.5. Tukey's fences identify outliers as values that fall beyond 1.5 times the interquartile range (IQR) above the

| ES | equivalent ES | valid range | Conversion formula | Source | correction factor |
|---|---|---|---|---|---|
| r | $\omega$, $\eta$ | -1 to 1 | - | - | - |
| Cohen's d | | $-\infty$ to $+\infty$ | $r \approx d/(d^2+4)^{0.5}$ | [53] | 1.27 |
| odds ratio (OR) | | 0 to + $\infty$ | $r \approx (OR^{0.5}-1)/(OR^{0.5}+1)$ | [52] | 1.46 |
| Kendall's $\tau$ | | -1 to 1 | $r \approx Sin(0.5 \times \pi \times \tau)$ | [71] | 1.01 |
| Spearman's $\rho$ | | -1 to 1 | $r \approx 2 \times Sin(\rho \times \pi/6)$ | [118] | 1.02 |
| $\eta^2$ | $R^2$, $\omega^2$, $\epsilon^2$ | 0 to 1 | $r \approx \sqrt{\eta^2}$ | [51, 53] | - |
| Cramer's V | | 0 to 1 | $r \approx V$ | [33] | 1.44 |
| CLES | | 0 to 1 | $r \approx sin((CLES-0.5) \times \pi)$ | [43] | - |
| Cohen's f | | 0 to $+\infty$ | $d \approx 2f$ - use conversion for Cohen's d | [31] | - |
| Cohen's $f^2$ | | 0 to $+\infty$ | $r \approx \sqrt{f^2/(f^2+1)}$ | [31] | - |

**Table 1: Overview over the effect size measures included in the meta-study, including valid ranges, formulas for conversion of effect sizes to r and sources for the conversion formulas. Correction factors are mean correction factors taken from Poom and af Wåhlberg [108]. Correction factors were not available for all effect size measures.**

75% quantile or below the 25% quantile [140]. We used the sample size as weights, since larger sample sizes lead to more reliable estimates [45] and sample sizes were more consistently reported and extracted than for example confidence intervals, which also help judge reliability. Calculating a per-paper median first avoids undue influence of a single paper with an extreme amount of tests on the average overall [45]. We did this separately for the within-groups and between-groups designs. We grouped these per-paper measures of central tendency according to the categories assigned to each paper to get the distributions of effect sizes for different research areas at CHI. We followed prior work on domain-specific effect size guidelines, which approximated small, medium and large effects in a specific research area, by dividing the distribution into thirds and calculating the median effect size within these thirds as a threshold between size categories [65, 91]. To counter the bias towards large effect sizes [45], we again used medians weighted by sample size.

## 4.5 Limitations

Even though there has been criticism of NHST and prior work has argued that Bayesian statistics are better suited for CHI's needs of gaining precise knowledge and estimates even from small sample studies [70], we focused our meta-study on NHST statistics and effect sizes, as this is still the most commonly referenced statistical paradigm at CHI. While Bayesian statistics may be more suitable, they also require a significant learning curve to get started [104]. Phelan et al. have investigated providing analysis templates to make it easier, but it was still hard for researchers without a statistics background to trust the priors recommended by the templates [104]. This means that wide-spread adoption of Bayesian statistics could still be some way off, so we believe it is worthwhile aiding knowledge accrual when NHST is used in the meantime.

Our LLM extraction process, like any use of LLM, can include hallucinations [61] or other mistakes. We did our best to mitigate their influence as much as possible. We used the lowest possible temperature settings to lower randomness in our extraction process. Additionally, we used automatic and manual plausibility checks especially on numeric values to check whether they actually exist in the papers. For the effect sizes in our meta-study, we reviewed that the effect size measure fit with the test it was reported with and additionally screened out values not in the valid range of values for the specific effect size measures. On the whole we designed our approach to err on the side of caution, rather accepting that effect sizes were missed than incorrect ones were included.

Our effect size meta-analysis is naturally only based on those tests which report effect sizes. Unfortunately these are not always reported. Additional information such as exact test/experimental design and sample sizes also are not always clear and we had to make inferences. Our ground truth analysis highlighted that our LLM missed effect sizes not clearly referenced as such in the text. Thus, our meta-analysis only provides a view on a subset of effects at CHI. Our LLM also did not always correctly identify the correct subsample N, leading to possible overemphasis of certain effects. Finally, the LLM did not always extract the correct test names, leading to uncertainty, regarding the between or within groups study designs. In practice, researchers also often interpret effect sizes from within and between groups designs in the same way [122]. However, despite these issues, since all extracted effect sizes in our ground-truth evaluation were correct, we believe our data set is large and robust enough to serve as a valuable starting point for the community. We believe this is a useful step forward over the state of the art. In future work, we hope to establish automatically generated and machine readable supplemental material containing all necessary information on tests to remove the above issues.

Converting effect sizes to a different effect size measure can also introduce biases. Poom and af Wåhlberg showed in their simulations that common conversion formula may lead to systematic misestimations [108]. While some effect sizes are derived from the same basis, they are not all numerically equivalent - being based on different experimental designs [31]. Wherever possible, we use the average correction factors provided in their work, but they were not available for all effect sizes we incorporated [108]. We nevertheless wanted to include as many effect sizes as possible, since selective exclusion may also bias the results of the meta-study.

For the most part, effect sizes in our analysis were sample based effect sizes, that aim to provide an estimate of the effect in the population. However, they can be biased estimators [75, 113]. Some less biased effect sizes measures, like $\omega^2$ for ANOVA-Type analyses

or Hedge's g for mean comparisons are recommended [75], and were also used to a small extent in our sample, but for our overview of effect sizes at CHI, both less and more biased effect size measures were incorporated, to the extent that we identified conversion formulas for them. When considering only published results, as we do in our work, publication bias has an influence on the size of reported effects [58]. Papers with statistically significant results are more likely to be published, but papers with small sample sizes (such as many of the studies in our sample) have poor power to detect small effects, so it is more likely that published effects are larger than true population effect sizes [45]. We fully expect that updates to our recommendations will be necessary as the CHI community moves towards more detailed effect size reporting practices. However, to provide assistance to researchers now, we have to make use of the data we have now, biased as it may be.

Finally, because the CCS we used to categorize our sample are used by the whole ACM, they are not specific to CHI. As such some of the discussed research areas, like *security and privacy* are more granular and useful than other more general ones, like *empirical studies in hci*. To enable researchers to make usable inferences for the categories they view as most interesting, we have published our dataset, including the extracted statistics and categorization per paper on OSF. Through matching by doi, researchers with access to different categorization metrics could include these.

## 4.6 Results

We present an overview of the distribution of effect sizes reported in different research areas at CHI. Our sample for the guidelines consists of 599 quantitative CHI papers from 2019 to 2023, which reported effect sizes we could clearly identify and convert to Pearson's r. The most frequently reported tests were ANOVA (18.0%), followed by repeated-measures ANOVA (14.5%) and Pearson's correlation (5.1%). Correspondingly the most frequently reported effect sizes were $\eta_p^2$ (21.3%), Pearson's r (16.7%), and $\eta^2$ without any specifiers (14.4%), followed by Cohen's d (12.7%) and OR (10.8%). Among all of the valid effect sizes, 6.4% were reported without any hypothesis test specified.

Figure 2 shows the distribution of effect sizes calculated over all 599 quantitative papers (incorporating 4381 tests and 6712 effect sizes) as well as for the individual research areas. We also differentiate between within-group and between-group experiment designs. In Table 2 we report the median sample sizes and the number of papers[3] used to calculate our thresholds. The higher these numbers are, the more robust the thresholds will likely be for a given research area. In our estimation, more caution is advised for the areas of interaction design, mixed / augmented reality, social and professional topics and computing methodologies. While we still find the thresholds useful for putting results in context, since it is the best context we have, for power analysis, we recommend adding some buffer because effect sizes might be inflated and/or unduly influenced by a small number of outliers.

Guidelines for all effect size measures included in our meta analysis can be found in Table 3. The research area specific size guidelines are provided in Table 4 for Pearson's r. We calculated the thresholds using Pearson's r, but for convenience, we provide the size guidelines converted for the other effect size measures, $\eta^2$, Cohen's d, OR ,CLES, Cohen's f, Cramer's V, $f^2$, Kendall's $\tau$ and Spearman's $\rho$, in the appendix.

According to Cohen's guidelines, r=0.1, 0.3, and 0.5 would be considered small, medium and large effects respectively [31]. For the overall between group effects in our CHI sample, these thresholds would be 0.14, 0.27, and 0.61 respectively, and for within group effects, 0.12, 0.32 and 0.72.

While the overall small and medium effect threshold for between groups at CHI is relatively similar to Cohen's guidelines, the threshold for large effects is higher. This is even more pronounced for within groups effects. But most importantly we see clear variation between the different research areas, e.g. with effect sizes generally lower in *security and privacy*, compared to *collaborative and social computing*.

Since sample size influences the size of effect that is detectable by a significance test, we also plotted the sample size distributions in Figure 3. We used both box plots to show summary statistics, and violin plots to give more information about the shape of the distribution. Plots display the sample sizes as used for our meta-analysis of effect sizes, e.g. the maximum sample size considered was 352.5, the cut-off for outliers according to Tukey [140] and all larger sample sizes were interpolated as this maximum sample size. This is visible in the slight overemphasis of distributions at that sample size. For most research areas, sample sizes are skewed towards small sample sizes, and are smaller for within groups designs, which makes sense since these have more power to detect effects, so fewer participants are needed. For *collaboration and social computing*, *security and privacy*, and *social and professional topics*, this skew is not as obvious and sample sizes are evenly spread or even skewed towards higher sample sizes. *Mixed/augmented reality* and *virtual reality* are also somewhat special cases, in that they exhibit comparatively lower sample sizes than other research areas, likely due to hardware and special equipment needed to conduct such studies, representing a limit on the number of participants it is possible to recruit.

Combining insights from both effect size and sample size distributions, it seems that effect sizes in some fields, like *security and privacy* are smaller than in others, even though these fields have access to comparatively large sample sizes, which would enable them to detect both small and large results. However, there is no clear pattern based on sample sizes, given that the effect size thresholds for *mixed/augmented reality*, which has primarily smaller sample sizes, are not higher than for the rest of CHI. Figure 4 shows the relationship of sample size and effect size for all papers included in the meta-study. There are negative correlations which would be judged as between medium and small in the context of all analyzed CHI papers, see Table 3. This may be because in some papers, effect sizes were only reported for significant results, so small effect sizes were possibly under-reported, effect sizes from small studies might be inflated, and it could also be an indicator for publication bias. However, around the median sample size (50 for between-groups and 28.5 for within-groups designs) effect sizes are distributed across the whole spectrum from very small to very large. In comparison, an analysis of 447 papers in psychology found

[3]Papers can contain both within and between group effects thus the numbers do not add up to 599

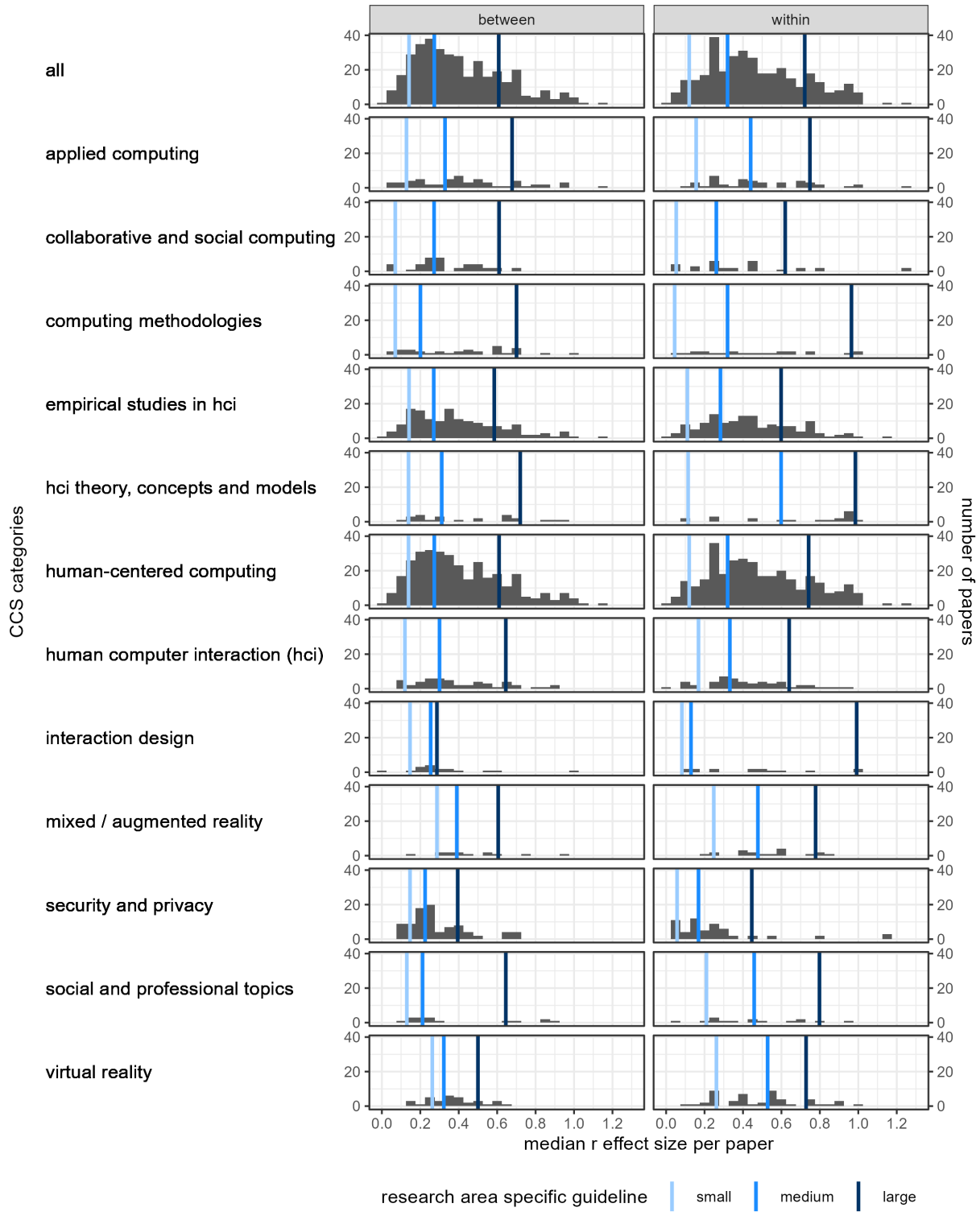


**Figure 2: Histograms of median effect sizes (Pearson's r) per paper, reported separately for different research areas and experiment designs. Research area specific size guidelines superimposed on the histograms.**

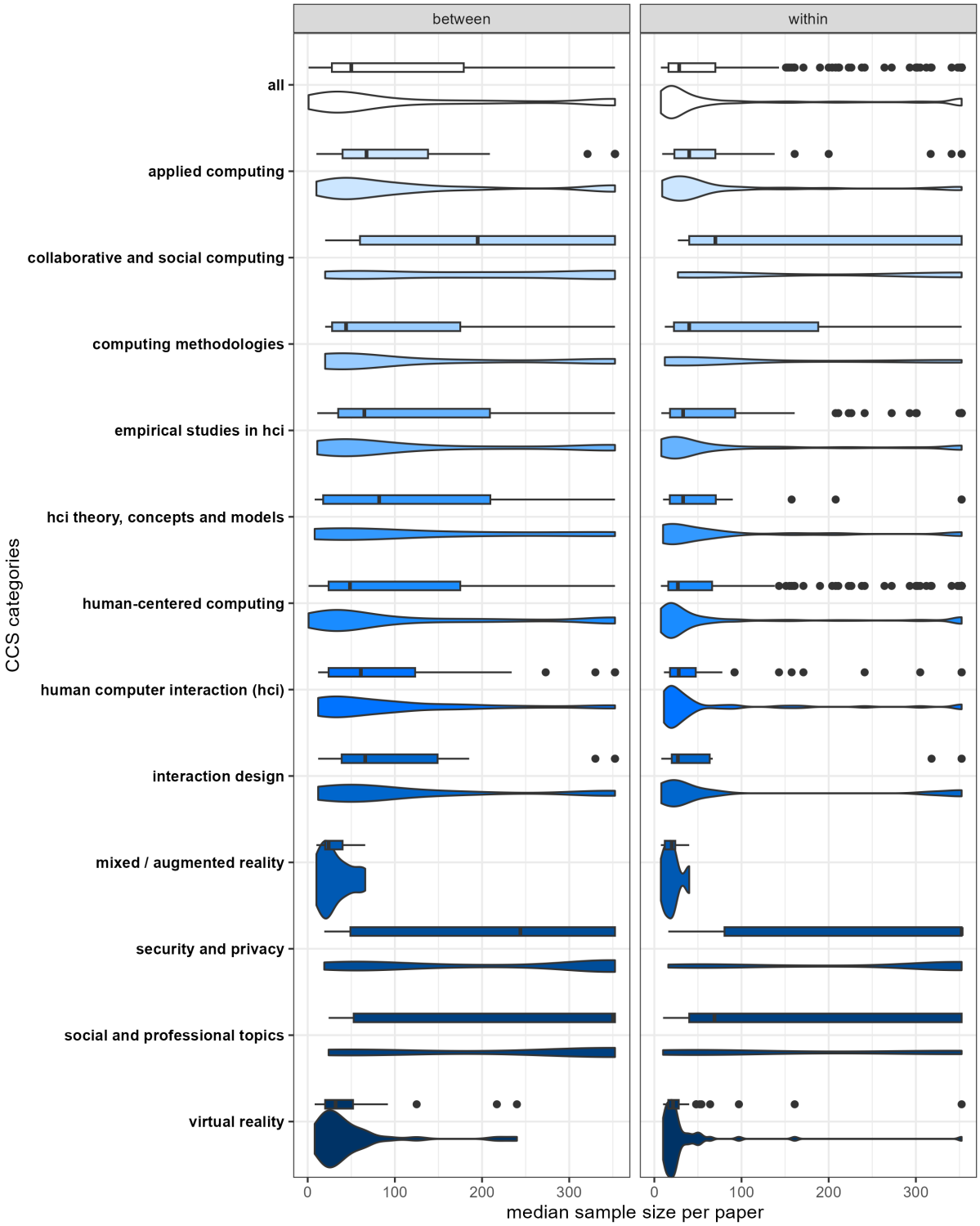


**Figure 3: Violin plots and box plots displaying the distribution of median sample sizes per paper, as used in the meta-study, separately for different research areas and experiment designs.**

| research area | between groups | | | | within groups | | | |
|---|---|---|---|---|---|---|---|---|
| | N° Paper | sample size | | | N° Paper | sample size | | |
| | | min | median | max | | min | median | max |
| all | 359 | 1 | 50.0 | 289 mill. | 346 | 4 | 28.5 | 7010271 |
| applied computing | 57 | 3 | 67.5 | 289 mill. | 43 | 6 | 40.0 | 100000 |
| collaborative and social computing | 41 | 20 | 195.0 | 289 mill. | 30 | 27 | 70.0 | 7010271 |
| computing methodologies | 33 | 20 | 44.0 | 2871 | 23 | 11 | 40.0 | 12000 |
| empirical studies in hci | 155 | 5 | 65.0 | 289 mill. | 138 | 6 | 33.0 | 65903 |
| hci theory, concepts and models | 24 | 8 | 82.0 | 1690 | 20 | 10 | 33.0 | 1132 |
| human computer interaction (hci) | 50 | 12 | 61.0 | 1690 | 51 | 9 | 28.0 | 65903 |
| human-centered computing | 332 | 1 | 48.5 | 289 mill. | 322 | 4 | 27.0 | 7010271 |
| interaction design | 17 | 12 | 66.0 | 1506 | 15 | 8 | 27.0 | 6690 |
| mixed / augmented reality | 13 | 10 | 24.0 | 66 | 18 | 7 | 20.0 | 40 |
| security and privacy | 89 | 19 | 244.0 | 11953 | 58 | 16 | 437.0 | 7010271 |
| social and professional topics | 16 | 24 | 577.5 | 26174 | 15 | 10 | 69.0 | 3155 |
| virtual reality | 33 | 8 | 32.0 | 240 | 50 | 9 | 21.5 | 9860 |

**Table 2: Median and range for sample sizes and number of papers per research area in the meta-study. Papers categorized as multiple research areas are included for all that are applicable.**

| effectsize measure | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| r, $\omega$, $\eta$ | 0.14 | 0.27 | 0.61 | 0.12 | 0.32 | 0.72 |
| $\eta^2$, $R^2$, $\omega^2$, $\epsilon^2$ | 0.02 | 0.07 | 0.37 | 0.01 | 0.10 | 0.52 |
| Cohen's d | 0.22 | 0.44 | 1.09 | 0.19 | 0.52 | 1.38 |
| OR | 1.48 | 2.13 | 5.90 | 1.39 | 2.44 | 8.72 |
| CLES | 0.54 | 0.59 | 0.71 | 0.54 | 0.60 | 0.76 |
| Cohen's f | 0.11 | 0.22 | 0.55 | 0.10 | 0.26 | 0.69 |
| Cramer's V | 0.10 | 0.19 | 0.42 | 0.08 | 0.22 | 0.50 |
| $f^2$ | 0.02 | 0.08 | 0.59 | 0.01 | 0.11 | 1.08 |
| Kendall's $\tau$ | 0.09 | 0.17 | 0.41 | 0.08 | 0.20 | 0.51 |
| Spearman's $\rho$ | 0.13 | 0.26 | 0.58 | 0.11 | 0.30 | 0.69 |

**Table 3: Guidelines including all research areas and for all effect size measures. The number of included papers is 359 for between-groups designs and 346 for within-groups designs. The median sample size is 50 for between-groups designs and 28.5 for within-groups designs.**

a correlation of r=-0.48, $CI_{95\%}$=[-0.56, -0.37] [4]. When compared to the results in Figure 4, the negative correlation in HCI is noticeably weaker than in psychology [74], such that the CIs hardly overlap.

For use in a-priori power analysis, these guidelines should only be a starting point for researchers' estimation. Researchers in areas where large effect sizes are common and thus small samples are sufficient, should consider whether detecting smaller effect sizes might also be of practical relevance. Researchers in areas where small effect sizes are common should not be discouraged from conducting studies with small sample sizes if they believe the effect they aim to detect is large, but of course if possible go for larger samples.

[4]The paper originally reported a Spearman correlation of $r_S$=-0.45, $CI_{95\%}$=[-0.53, -0.35], which we converted using the formula in Table 1.

## 5 Effect Size Interpretation

While our research area specific guidelines can help researchers interpret their results in the context of typical effect sizes of their area better than Cohen's guidelines, we do not suggest using them blindly. They should serve as context but further qualitative interpretation is very valuable.

We conducted a qualitative analysis of effect size interpretation in quantitative CHI papers published 2023 and identified five categories of approaches employed by authors to interpret effects and effect sizes. Currently, effect sizes are rarely used to interpret results in depth. The most common form of effect discussion is the use of simple comparisons of descriptive statistics. In the rare cases of a more qualitative interpretation e.g. a discussion of practical relevance, the incorporation of effect size values is even less frequent.

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.14 | 0.27 | 0.61 | 0.12 | 0.32 | 0.72 |
| applied computing | 0.13 | 0.33 | 0.68 | 0.16 | 0.44 | 0.75 |
| collaborative and social computing | 0.07 | 0.27 | 0.61 | 0.05 | 0.26 | 0.62 |
| computing methodologies | 0.07 | 0.20 | 0.70 | 0.04 | 0.32 | 0.96 |
| empirical studies in hci | 0.14 | 0.27 | 0.58 | 0.11 | 0.28 | 0.60 |
| hci theory, concepts and models | 0.14 | 0.31 | 0.72 | 0.12 | 0.60 | 0.98 |
| human computer interaction (hci) | 0.12 | 0.30 | 0.64 | 0.17 | 0.33 | 0.64 |
| human-centered computing | 0.14 | 0.27 | 0.61 | 0.12 | 0.32 | 0.74 |
| interaction design | 0.15 | 0.25 | 0.29 | 0.08 | 0.13 | 0.99 |
| mixed / augmented reality | 0.29 | 0.39 | 0.60 | 0.25 | 0.48 | 0.78 |
| security and privacy | 0.15 | 0.22 | 0.40 | 0.06 | 0.17 | 0.45 |
| social and professional topics | 0.13 | 0.21 | 0.64 | 0.21 | 0.46 | 0.80 |
| virtual reality | 0.26 | 0.32 | 0.50 | 0.26 | 0.53 | 0.73 |

**Table 4: Guidelines for separate research areas, using Pearson's r**

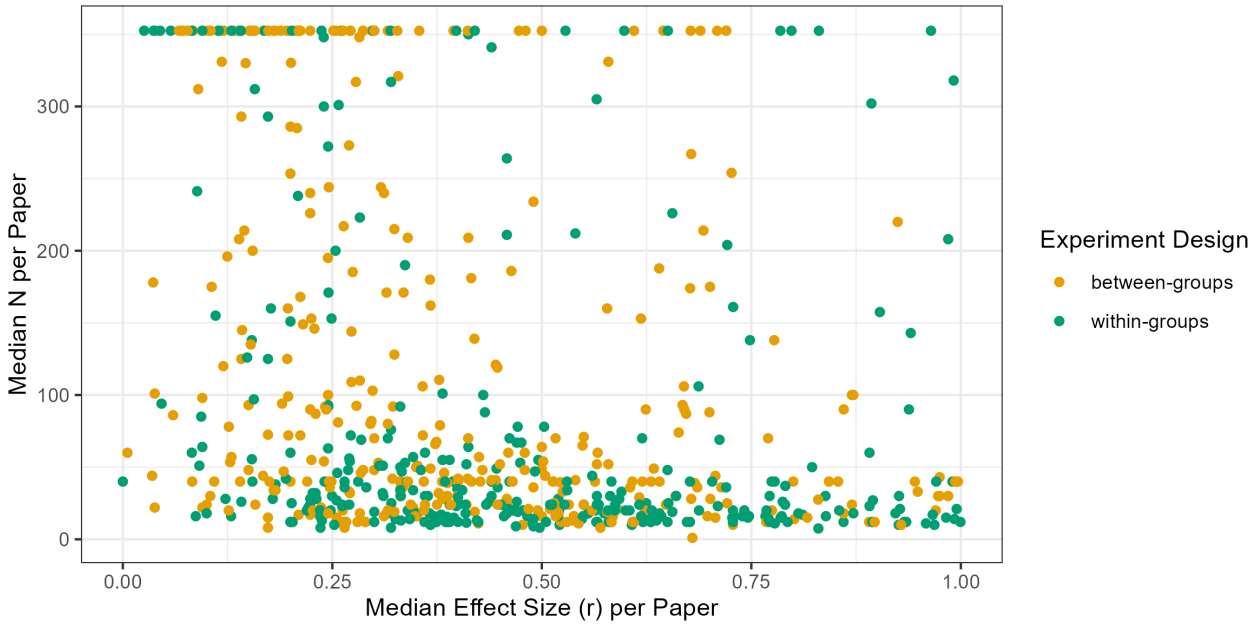


**Figure 4: Effect size plotted against sample size per paper and experiment design. For between-groups designs, the Pearson correlation was r=-0.29, $CI_{95\%}$=[-0.38, -0.19], and for within-groups designs, it was r=-0.24, $CI_{95\%}$=[-0.34, -0.14]. Outliers in sample size are plotted at 352.5 (value used as maximum weight in meta-study).**

## 5.1 Sample Selection

For our qualitative analysis, we chose CHI'23 as the most recently published at the start of our analysis. We randomly sampled quantitative papers as categorized by our filtering (see Section 3.1). We employed saturation regarding the interpretation approaches as a stopping criterion, i.e. we continued sampling as long as we were identifying new approaches. We coded all approaches we found, even if they only occurred once. We stopped after not identifying any new approaches in 5 papers. We then analyzed 5 additional papers to validate the approaches we had identified. Three papers that contained statistics but were *largely qualitative papers* were excluded from this process. This resulted in 67 analyzed papers. We provide a list of our sample in the supplemental material.

## 5.2 Data Analysis

In our coding process, we focused on two main goals. We refer to the first goal as *reporting coding*. In this step of our process, we

used deductive coding to answer the following questions for the papers we analyzed:

- Are (standardized) effect sizes reported with hypothesis tests?
- If yes: Which effect sizes are reported?
- Is there a size judgment reported with the effect size?

Two coders (C1 and C2) established a codebook based on these research questions and an initial coding of 15 papers. We coded p-values, standardized effect sizes and size judgments for the papers. We considered this simple coding, as there are clear true and false answers [95]. As such, after establishing a shared codebook for this aspect of the coding process, a single researcher analyzed each paper. If they were unsure about something, they discussed it with the other researchers.

We refer to our second goal as *interpretation coding*. In this step, we used inductive coding to answer the question

- How are effect sizes interpreted?

C1 and C2 established an initial codebook based on the same 15 papers. They and one additional coder C3 then coded individually. We consider this aspect to be complex coding, as it leaves more room for interpretation [95], so C1 and C3 went over all the coded instances of interpretations, verified them and grouped them to identify approaches. We set a rather low threshold for what we considered interpretation. Authors had to go beyond merely reporting the standardized effect size as is, or alongside descriptive statistics. Any further elaboration, e.g. specifying the direction of an effect or rephrasing to explain the effect size, was considered interpretation.

We additionally collected the following information about each paper:

**Type of paper** Within the papers considered quantitative by our filters, we identified four flavors: *NHST-quantitative*, *method*, *largely qualitative* and *Bayesian*.

**Has descriptive statistics** For each paper we noted whether it reported descriptive statistics with at least one hypothesis test. We considered descriptive statistics to be e.g. means, standard deviations or occurrence counts, depending on the hypothesis test.

## 5.3 Limitations

Our qualitative evaluation of effect size interpretation is limited by the expertise of the analyzing authors in the various research areas at CHI. While we use NHST in our own research and thus can interpret statistical results, due to our random selection of papers from across research areas, we are not equally familiar with the common measures, terminology and goals of each research area. When analyzing interpretation approaches, the extent of interpretation necessary to arrive at a conclusion was thus not always straightforward. However, we derived approaches to interpretation that were applicable across research domains.

## 5.4 Types of Papers

During our qualitative analysis, we identified four types of paper within our sample. *NHST-quantitative* papers were the expected default. These papers reported quantitative or mixed-methods results for HCI related research questions. They contained a substantial amount of hypothesis testing following the NHST paradigm. We applied both reporting and *interpretation coding* to these papers. *Method* papers developed new methods in HCI, such as development or validation of new measurement scales or discussing of analysis methods, without applying them to a content-related research question. *Bayesian* papers used only or mainly Bayesian statistics in reporting the results for content-related research questions. As we focus our contributions on providing guidance for choosing effect sizes for power analysis and interpreting results in HCI, i.e. for content-related studies in contrast to *method papers*, and as *Bayesian* papers report different measures of effect sizes, we only applied the *interpretation coding* to these papers and did not apply the *reporting coding*. *Largely qualitative* papers only report hypothesis tests which are not relevant for answering the main content-related research questions, e.g. Jacobsen et al. [60] employed two tests to detect shifts in interaction frequency across households, but otherwise focused on findings from their qualitative analysis. The tests were not discussed further, and we did not consider these papers in our analysis.

## 5.5 Overview of Effect Size Reporting

We analyzed a total of 67 papers. Of these, we categorized 57 as *NHST-quantitative*, 6 and 1 as *method* and *Bayesian* papers respectively and 3 as *largely qualitative*. Of the 57 *NHST-quantitative* papers, 30 reported at least one standardized effect size, and 10 reported effect sizes accompanied by a size judgment at least once. 2 of the papers pointed out a source for their size judgments. 42 of 57 *NHST-quantitative* papers contained descriptive statistics alongside at least one test.

## 5.6 Interpretation Approaches

We identified five categories of approaches used by CHI'23 authors in interpreting effects and effect sizes in their work. We situate them in a spectrum from descriptive approaches, which provide additional, often numerical and objective information to interpret effect sizes, to interpretative approaches, which include in-depth reasoning beyond the immediate scope of the work in question. Figure 5 shows this spectrum, while Table 5 provides specific examples from our sample of analyzed papers for each of the five categories. In the following, we describe these approaches in more detail.

On our spectrum from descriptive to interpretative approaches to interpreting effect sizes, the most descriptive category is *comparing values*, which are part of the output from the statistical analysis. The simplest is stating the *direction of the effect*, e.g. when comparing groups, which group was faster or made less mistakes. Reporting the effect size *in units of the study* is a bit more involved than just reporting descriptive statistics, i.e. reporting the difference in group means in the units in which it was measured instead of just the group means. This is a form of non-standardized, or simple effect size. Differences can also be compared *in relative units*, using phrases like "20% faster". This abstracts away the units of the simple effect size, but is still close to the measurement methods of the study. However, using such relative units, e.g. percentages, can be confusing as well, as Dragicevic has elaborated specifically for reporting of speed differences [41]. He recommends reporting ratios or percent differences if relative comparisons are required [41]. If standardized effect sizes are reported, then *comparing effect sizes*

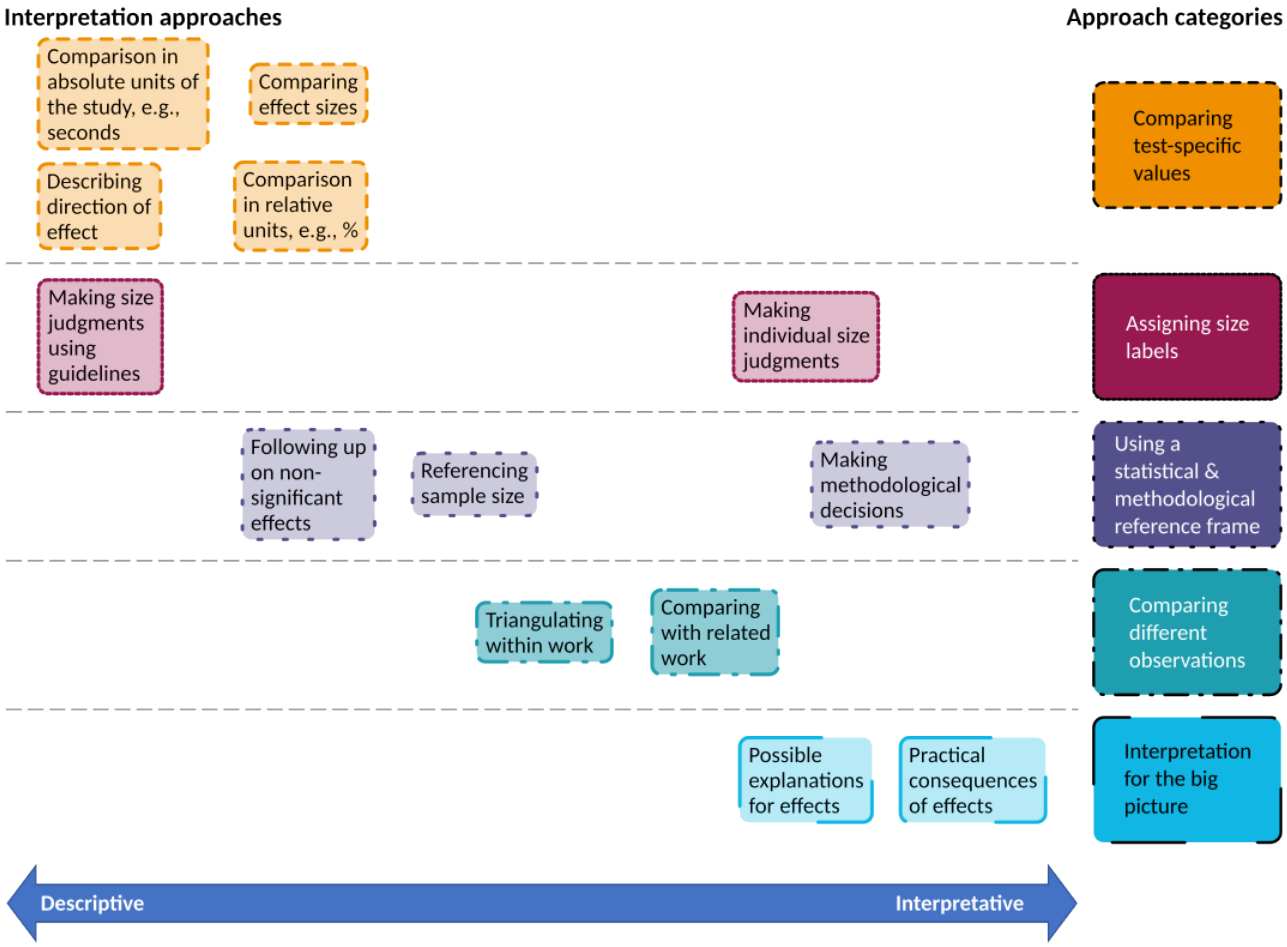


**Figure 5: Visualization of the spectrum from descriptive to interpretative interpretation approaches. We display approaches and their respective category on a single horizontal level each. The distance to the x-axis does not carry meaning otherwise.**

within a study, or to a set threshold, is another approach we saw. In this approach, the effect sizes themselves are included in the comparison, rather than descriptive statistics.

*Assigning labels* can be considered descriptive or interpretative, depending on how researchers address them. 'Labels' in this context refer to size judgments, e.g., researchers denoting an effect as having a specific size. *Making size judgments using guidelines* was the more common approach, using guidelines for determining a small, medium or large effect, such as those by Cohen [31], even if these guidelines were not always explicitly named. This approach falls on the descriptive end of the spectrum. However, when *making individual size judgments*, the judgments are not obviously based on published guidelines. If they are explained in the context of the study, we deem these judgments to be interpretative.

Some researchers are *using a statistical and methodological reference frame*, i.e., they reference statistical or methodological decisions or consequences in their interpretation of effect sizes. When *following up on non-significant effects* researchers use other descriptive approaches of interpreting effect sizes to get around the binary notion of statistical significance and highlight differences in their results despite non-significant tests. Researchers *referencing sample size* interpreted effect size in the light of the sample size of their study, especially relating to under-powered or over-powered hypothesis tests. Finally, effect sizes were used to *make methodological decisions*, e.g., regarding model choice, predictor selection and merging of data sets within a study, which we consider more interpretative.

*Comparing different observations* involves comparing results from a different data set, in contrast to comparing individual values within a data set. This can happen as *triangulation within work*, where other data sources presented in the same study are used to back up quantitative results. In our sample, these sources were mainly qualitative or descriptive. Another approach is *comparing with related work*, where the comparison is outside the study currently being reported, with published related work. Both comparison approaches can show similarities, where the different observations support the current result, or comprise of contradictory results. We allocate these approaches to the middle of the spectrum between descriptive and interpretative approaches. Comparisons

| Approach category | Approach | Example Quote |
|---|---|---|
| Comparing test-specific values | Direction of effect | "However, writing with multiple suggestions leads to slightly longer texts." [35] |
| | In absolute units of the study | "[...] significantly increased the number of words in a sentence after which suggestions were requested – **by about 1.5 words** (Means: s =10.93, s =9.48; i.e. a relative increase of 15.3%)" [35] [emphasis added] |
| | In relative units | "[...] significantly increased the number of words in a sentence after which suggestions were requested – by about 1.5 words (Means: s =10.93, s =9.48; **i.e. a relative increase of 15.3%)**" [35] [emphasis added] |
| | Comparing effect sizes | "Also, the correlation coefficient between profile type and recommendation ratings was higher (and sometimes statistically significant) for museum users." [154] |
| Assigning size labels | Making size judgments using guidelines | "Both plank (static core) and press-ups (dynamic upper) showed only non-significant, small effect sizes (Cohen's d = 0.239 and 0.298)." [29] |
| | Making individual size judgments | "These differences are rather small, about 6-10 words" [35] |
| Using a statistical and methodological reference frame | Following up on non-significant effects | "Finally, it is worth nothing that in both cases the effect is in the right direction ( F > P ), however the effect sizes are small, meaning that the study would have been under-powered to detect such small changes" [29] |
| | Referencing sample size | "However, since our sample size is relatively small (N=24 after two removed samples), there are limitations regarding the certainty of this observation" [124] |
| | Making methodological decisions | "This means that the model is able to explain 99.98% of the shared variance between the Hexad-12 and the Hexad-24. This result clearly shows that the two variable sets are not independent. Thus, we can continue analysing the results of the dimension reduction analysis to check whether the predictor variables load on the same CF as the criterion variables." [73] |
| Comparing different observations | Comparing with related work | "Our findings replicate [41] that individual mood does not change much with effective peer counseling, but contrasts with those of [3]" [149] |
| | Triangulating within work | "for concurrent joint attention, the charts area of interest (Figure 6a) all tree conditions perform equally well [sic]. This result is consistent with the Q6 question of the questionnaire, which indicates no statistical difference across the experimental conditions. Qualitative analysis of interviews suggests that this result could be due to the effectiveness of verbal communication in refining and specifying the area of interest " [16] |
| Interpretation for the big picture | Possible explanations for effects | "We believe there are two potential explanations. First, our study procedure is straightforward. An improved consent form reading may not have a strong effect in preparing the participants for the later study. Second, our study is low-stake compared to medical trials." [152] |
| | Practical consequences or limitations of effects | "Although our MR system helped to improve the success rate and precision of the deformation, 4Doodle has limited accuracy compared to machine printing because of the uncertainty associated with human performance." [132] |

**Table 5: Examples of interpretation approaches. All citations within the quotes refer to the bibliography of the original paper.**

can reproduce binary judgments about outcomes, e.g. if an intervention is effective or not, in which case they are more descriptive, or they can go into more interpretative depth regarding connections between the observations and comparing actual sizes of effects.

On the interpretative side of the spectrum, researchers provide *interpretation for the big picture*, i.e. beyond the current study, as context around effects and effect sizes. They speculate on *possible explanations for effects*, e.g. based on other statistical results, common knowledge, or prior work. Researchers also discuss *practical consequences of effects* for further research and theory, but also deployment of newly developed techniques in practice or in real-world scenarios.

In considering which approach to use, we do not believe that one category or one side of the spectrum is better than the other. Instead, they are suitable for different use cases. The more descriptive approaches in our spectrum are easier to implement, as values to compare may be provided automatically from statistical analysis software, or can be calculated based on existing data. As such these approaches can be used for all or most of the statistical tests reported in a paper. Since descriptive approaches refer to a single concrete effect, they should be reported close the hypothesis tests.

More interpretative approaches require more effort and knowledge of a field and are thus more dependent on individual researchers' backgrounds and judgments, in addition to needing more space to describe implications. They often bring up prior

work and implications for the future or practicality of deployment to assess the relevance of not just a single effect size, but multiple test results taken together for a big picture. Anvari et al. provide a thorough summary of influencing factors on practical relevance with examples from psychological research [2]. When judging practical consequences of effects, even small effects can be immensely beneficial if many people are affected [2] and no larger effect interventions exist or are practical. On the other hand, the importance of large effects can be diminished through practicality or interactions with other effects or not being as generalizable as anticipated [2]. This is especially important for the main hypotheses and contributions of a paper. Interpretative approaches fit in the discussion, where a broader context can be taken into account. However, to maintain a grounding of the interpretation in the statistical results, numbers from effect size calculations should be referenced in such discussions, instead of a binary statement that an effect exists or not based on statistical significance.

In addition, not all approaches are equally applicable to all types of research. This applies especially to the category *Using a statistical & methodological reference frame*, wherein e.g. effect sizes can only be used to *make methodological decisions*, if testing was conducted for this purpose and if results are statistically significant, then *following up on non-significant effects* is not necessary. When deciding which approach to use, researchers should thus take into account one that is suitable for their goal.

## 6 Recommendations

Based on our meta-study and qualitative analysis, we have the following recommendations on improving reporting of statistical results, especially effect sizes.

**Report all effect sizes, including those which are not significant.** To gain a full understanding of the research area, non-significant effects are also important to report. To avoid overloading the results section with numbers, consider reporting the full statistics in an appendix.

**Interpret effect sizes within your research context.** As our meta-study showed, effect sizes can vary strongly between different research areas and thus it makes sense to compare a found effect size with what is common in the area. For the effect sizes r, Cohen's d, OR, $\omega$, $\omega^2$, $\eta$, $\eta^2$, $\epsilon^2$ and CLES, see the Tables 4 - 9 from this paper, and for the effect sizes Kendall's $\tau$, Spearman's $\rho$, Cramer's V, Cohen's f and Cohen's $f^2$, see the tables in the supplemental material.

**Discuss effects on an appropriate level of detail based on relevance.** Differentiate between the important effect sizes which need to be discussed in detail, e.g. the main hypothesis tests, compared to a potentially large number of secondary tests, e.g. pairwise post-hoc tests. Researcher discretion is needed.

**Interpret effect sizes qualitatively based on practical relevance.** Apart from stating the size of the effect in the context of the research area, for relevant tests, also discuss the practical relevance. Our effect size guidelines are meant as a help to authors, reviewers and readers for judging the relative size of an effect. We want to make absolutely clear that they should not be used to arbitrarily accept or reject research work and that small effects can be highly practically relevant and large effects practically irrelevant.

In the following we describe how the guidelines and findings in our work can be used by HCI researchers.

1) Researchers planning a study have the study design figured out and now need to decide how many participants they want to recruit.

2) They know which statistical tests they want to use to analyze the data, and which type of effect size to report with it. They can use various ways to decide on an appropriate sample size, including resource constraints, prior work or deciding on the minimal effect size of interest [77].

3) For this example, let's assume the researchers' topic is very novel, so there is no direct related work they can use to determine what effect size to expect.

4) Now there are two options. One, if based on their expert judgment, they have a minimum relevant effect they want to find, they can use our tables to check whether this effect size is in line with other effect sizes in the area of the planned study. Two, if the researchers do not have this intuition, they can simply use our tables to see what small, medium and large effect might be in their area. If the researchers' study domain is not specifically covered by the guidelines, e.g. *health*, they could instead look at a more general category, like *empirical studies in hci*. The idea is that researchers always take the most specific guideline as possible.

5) Should researchers see that the effect size they are targeting is smaller than what is typically found in their field, they need to carefully assess whether their experimental set-up is capable of detecting such small effects, e.g. by including mechanisms to reduce random errors or ensuring their sample size is large enough.

6) Based on this, the researchers can now conduct a power analysis, e.g. as explained for simple tests in the tutorials accompanying [96] [5] or for more complex designs by [19, 78].

7) Once the study is completed, and the authors have their effect sizes, they can use our tables to discuss them in the context of the related effect sizes in their field. While this step is the last step in the tutorial, we believe it is the most important step, because this significantly eases interpretation of results. Using different interpretation strategies highlighted in this paper, they can e.g. assign size labels to contextualize the effect size or discuss practical consequences of the effect they found on their population of interest.

Should researchers be conducting studies with different effect size measures, the table with the most similar effect size measure can be used as a starting point. E.g. for Hedge's g, Cohen's d values can be adjusted, given that Hedge's g is usually a bit lower than Cohen's d, since it is less biased [75]. However, this is less reliable than our empirical analysis for the effect sizes in this paper. Future work can look at explicitly extending our work to more effect size measures.

## 7 Conclusion

In this work, we extracted statistics from quantitative CHI papers published between 2019 and 2023 and used them to derive effect size guidelines specific to 12 research areas at CHI. Researchers can use these guidelines to interpret their results within their research area or as a starting point for a-priori power analysis. To further aid

[5] https://powerdb.info/tutorials

researchers in interpreting effect sizes, we analyzed effect size interpretation approaches in quantitative papers published at CHI'23 and identified five categories of approaches. We recommend that researchers provide additional descriptive interpretation for the majority of their reported hypothesis tests, and more in-depth qualitative interpretation including practical consequences of the main result.

## 8 Future Work

Our guidelines for interpreting effect sizes in the different CHI research areas need to be updated on a regular basis to reflect new developments in the field. To ease this process and make it more robust, we recommend that statistical libraries generate machine-readable and complete data of statistical tests, including subgroup sample sizes, effect sizes, confidence intervals, test names, test statistics and descriptive statistics, which can be published as supplemental material with little extra effort for the authors. With this future general meta-studies could be fully automated.

## Acknowledgments

The authors would like to thank Antonia Sistig and Lukas Wenz for their contributions in the development of the plausibility checker, Dorian Heidorn for extracting CCS information and Maximilian Mundt for his contribution to the coding of effect size interpretations This work was partially funded by the Werner Siemens Foundation.

## References

[1] Andrew D. Althouse, Jennifer E. Below, Brian L. Claggett, Nancy J. Cox, James A. de Lemos, Rahul C. Deo, Sue Duval, Rory Hachamovitch, Sanjay Kaul, Scott W. Keith, Eric Secemsky, Armando Teixeira-Pinto, Veronique L. Roger, and null null. 2021. Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association. *Circulation* 144, 4 (July 2021), e70–e91. https://doi.org/10.1161/CIRCULATIONAHA.121.055393

[2] Farid Anvari, Rogier Kievit, Daniël Lakens, Charlotte R Pennington, Andrew K Przybylski, Leo Tiokhin, Brenton M Wiernik, and Amy Orben. 2023. Not All Effects Are Indispensable: Psychological Science Requires Verifiable Lines of Reasoning for Whether an Effect Matters. *Perspectives on Psychological Science* 18, 2 (2023), 503–507. https://doi.org/10.1177/17456916221091565

[3] Farid Anvari and Daniël Lakens. 2021. Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. *Journal of Experimental Social Psychology* 96 (Sept. 2021), 104159. https://doi.org/10.1016/j.jesp.2021.104159

[4] American Psychological Association (Ed.). 2020. *Publication Manual of the American Psychological Association* (seventh edition ed.). American Psychological Association, Washington, DC.

[5] Donald R. Atkinson, Michael J. Furlong, and Bruce E. Wampold. 1982. Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is There a (Statistically) Significant Relationship? *Journal of Counseling Psychology* 29, 2 (1982), 189–194. https://doi.org/10.1037/0022-0167.29.2.189

[6] Thom Baguley. 2009. Standardized or Simple Effect Size: What Should Be Reported? *British Journal of Psychology* 100, 3 (2009), 603–617. https://doi.org/10.1348/000712608X377117

[7] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. 2021. Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3411764.3445584

[8] Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 33–39. https://doi.org/10.48550/arXiv.1904.08783 arXiv:1904.08783 [cs]

[9] Ac Bd, Christine Bauer, and Afsaneh Doryab. 2016. Solving the Battle of First-Authorship: Using Interactive Technology to Highlight Contributions. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, San Jose California USA, 609–620. https://doi.org/10.1145/2851581.2892582

[10] Brett A. Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. 2023. Programming Is Hard - or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (Sigcse 2023)*. Association for Computing Machinery, New York, NY, USA, 500–506. https://doi.org/10.1145/3545945.3569759

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[12] Donald D Bergh, Barton M Sharp, Herman Aguinis, and Ming Li. 2017. Is There a Credibility Crisis in Strategic Management Research? Evidence on the Reproducibility of Study Findings. *Strategic Organization* 15, 3 (Aug. 2017), 423–436. https://doi.org/10.1177/1476127017701076

[13] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290607.3310432

[14] Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune, and Antoine Doucet. 2022. Assessing the Impact of OCR Noise on Multilingual Event Detection over Digitised Documents. *International Journal on Digital Libraries* 23, 3 (Sept. 2022), 241–266. https://doi.org/10.1007/s00799-022-00325-2

[15] Frank A Bosco, Herman Aguinis, Kulraj Singh, James G Field, and Charles A Pierce. 2015. Correlational Effect Size Benchmarks. *Journal of Applied Psychology* 100, 2 (2015), 431.

[16] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2023. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 18. https://doi.org/10.1145/3544548.3581283

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Vancouver, Canada, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[18] Emeline Brulé, Brianna J. Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. 2020. Review of Quantitative Empirical Evaluations of Technology for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Chi '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376749

[19] Marc Brysbaert and Michaël Stevens. 2018. Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition* 1, 1 (Jan. 2018), 9. https://doi.org/10.5334/joc.10

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[21] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 981–992. https://doi.org/10.1145/2858036.2858498

[22] Paul Cairns. 2007. HCI... Not As It Should Be: Inferential Statistics in HCI Research. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. British Computer Society, Lancaster, UK, 7 pages. https://doi.org/10.14236/ewic/HCI2007.20

[23] Yiqun T. Chen, Angela D. R. Smith, Katharina Reinecke, and Alexandra To. 2023. Why, When, and from Whom: Considerations for Collecting and Reporting Race and Ethnicity Data in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)*. Association for Computing Machinery, New York, NY, USA, Article 395, 15 pages. https://doi.org/10.1145/3544548.3581122

[24] CHI 2025. 2024. Guide to a Successful Submission – CHI 2025. https://chi2025.acm.org/guide-to-a-successful-submission/.

[25] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody Speaks That Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Chi '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376569

[26] Defne Circi, Ghazal Khalighinejad, Shruti Badhwar, Bhuwan Dhingra, and L. Brinson. 2023. Retrieval of Synthesis Parameters of Polymer Nanocomposites

Using LLMs. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*. OpenReview.net, New Orleans, USA, 9 pages. https://openreview.net/forum?id=qpQr8px2Pz

[27] Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L. Catherine Brinson. 2024. How Well Do Large Language Models Understand Tables in Materials Science? *Integrating Materials and Manufacturing Innovation* 13, 3 (July 2024), 19 pages. https://doi.org/10.1007/s40192-024-00362-6

[28] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 3882–3890. https://doi.org/10.24963/ijcai.2020/537

[29] Christopher Clarke, Jingnan Xu, Ye Zhu, Karan Dharamshi, Harry McGill, Stephen Black, and Christof Lutteroth. 2023. FakeForward: Using Deepfake Technology for Feedforward Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581100

[30] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173715

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

[32] Kovila Coopamootoo and Thomas Gross. 2019. *A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies*. Technical Report CS_TR-1518. Newcastle University School of Computing.

[33] Harald Cramér. 1946. *Mathematical Methods of Statistics*. Princeton University Press.

[34] Sophia Crüwell, Johnny van Doorn, Alexander Etz, Matthew C. Makel, Hannah Moshontz, Jesse C. Niebaum, Amy Orben, Sam Parsons, and Michael Schulte-Mecklenbeck. 2019. Seven Easy Steps to Open Science. *Zeitschrift für Psychologie* 227, 4 (Oct. 2019), 237–248. https://doi.org/10.1027/2151-2604/a000387

[35] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models Using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3580969

[36] Yuntian Deng, David Rosenberg, and Gideon Mann. 2019. Challenges in End-to-End Neural Scientific Table Recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sydney, Australia, 894–901. https://doi.org/10.1109/ICDAR.2019.00148

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, Minneapolis, USA, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[38] Julian di Stephano. 2003. How Much Power Is Enough? Against the Development of an Arbitrary Convention for Statistical Power Calculations. *Functional Ecology* 17, 5 (2003), 707–709.

[39] Leonie Disch, Angela Fessl, and Viktoria Pammer-Schindler. 2022. Designing for Knowledge Construction to Facilitate the Uptake of Open Science: Laying out the Design Space. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3517450

[40] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[41] Pierre Dragicevic. 2016. *My Technique Is 20% Faster: Problems with Reports of Speed Improvements in HCI*. Report. Inria Saclay Ile de France, Inria Saclay, Ile de France.

[42] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300295

[43] William P. Dunlap. 1994. Generalizing the Common Language Effect Size Indicator to Bivariate Normal Correlations. *Psychological Bulletin* 116, 3 (1994), 509–511. https://doi.org/10.1037/0033-2909.116.3.509

[44] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. Touchstone2: An Interactive Environment for Exploring Trade-Offs in HCI Experiment Design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300447

[45] Paul D. Ellis. 2010. *The Essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK.

[46] EQUATOR Network. n.d.. Reporting Checklists for Medical Researchers. https://www.goodreports.org/.

[47] EQUATOR Network. n.d.. Reporting Guidelines. https://www.equator-network.org/reporting-guidelines/.

[48] 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 (May 2007), 175–191. https://doi.org/10.3758/BF03193146

[49] Christopher J Ferguson. 2009. An Effect Size Primer: A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice* 40, 5 (2009), 532–538. https://doi.org/10.1037/a0015808

[50] Piyum Fernando and Stacey Kuznetsov. 2020. OScH in the Wild: Dissemination of Open Science Hardware and Implications for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376659

[51] Andy Field, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using R* (1. edition ed.). SAGE Publications Ltd, London ; Thousand Oaks, Calif.

[52] Andy P. Field and Raphael Gillett. 2010. How to Do a Meta-analysis. *Brit. J. Math. Statist. Psych.* 63, 3 (2010), 665–694. https://doi.org/10.1348/000711010X502733

[53] Catherine O. Fritz, Peter E. Morris, and Jennifer J. Richler. 2012. Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of Experimental Psychology: General* 141, 1 (2012), 2–18. https://doi.org/10.1037/a0024338

[54] Thomas Groß. 2021. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. In *Socio-Technical Aspects in Security and Trust (Lecture Notes in Computer Science)*, Thomas Groß and Theo Tryfonas (Eds.). Springer International Publishing, Cham, 3–26. https://doi.org/10.1007/978-3-030-55958-8_1

[55] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. 2020. How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376454

[56] Beat Hulliger. 2023. *Modi: Multivariate Outlier Detection and Imputation for Incomplete Survey Data*. https://CRAN.R-project.org/package=modi R package version 0.1.1.

[57] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2, 8 (Aug. 2005), e124. https://doi.org/10.1371/journal.pmed.0020124

[58] John P. A. Ioannidis. 2008. Why Most Discovered True Associations Are Inflated. *Epidemiology* 19, 5 (2008), 640–648. jstor:25662607

[59] John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. 2015. Meta-Research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology* 13, 10 (Oct. 2015), e1002264. https://doi.org/10.1371/journal.pbio.1002264

[60] Rune Møberg Jacobsen, Kasper Fangel Skov, Stine S Johansen, Mikael B. Skov, and Jesper Kjeldskov. 2023. Living with Sound Zones: A Long-Term Field Study of Dynamic Sound Zones in a Domestic Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Association for Computing Machinery, New York, NY, USA, Article 405, 14 pages. https://doi.org/10.1145/3544548.3581535

[61] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (March 2023), 248:1–248:38. https://doi.org/10.1145/3571730

[62] AHA / ASA Journals. 2023. AHA / ASA - Statistical Reporting Recommendations. https://www.ahajournals.org/statistical-recommendations.

[63] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-Making under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300432

[64] Alex Kale, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. 2023. MetaExplorer : Facilitating Reasoning with Epistemic Uncertainty in Meta-analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. https://doi.org/10.1145/3544548.3580869

[65] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. 2007. A Systematic Review of Effect Size in Software Engineering Experiments. *Information and Software Technology* 49, 11 (2007), 1073–1086. https://doi.org/10.1016/j.infsof.2007.02.015

[66] Maurits Kaptein and Judy Robertson. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1105–1114. https://doi.org/10.1145/2207676.2208557

[67] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language

Models for Education. *Learning and Individual Differences* 103 (April 2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[68] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, San Jose California USA, 1081–1084. https://doi.org/10.1145/2851581.2886442

[69] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving Transparent Statistics Forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 534–541. https://doi.org/10.1145/3027063.3027084

[70] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 4521–4532. https://doi.org/10.1145/2858036.2858465

[71] Maurice George Kendall. 1970. *Rank Correlation Methods* (4 ed.). Griffin, Oxford, England.

[72] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting Scientific Results in Perspective: Improving the Communication of Standardized Effect Sizes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 625, 14 pages. https://doi.org/10.1145/3491102.3502053

[73] Jeanine Krath, Maximilian Altmeyer, Gustavo F. Tondello, and Lennart E. Nacke. 2023. Hexad-12: Developing and Validating a Short Version of the Gamification User Types Hexad Scale. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3580968

[74] Anton Kühberger, Astrid Fritz, and Thomas Scherndl. 2014. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE* 9, 9 (Sept. 2014), e105825. https://doi.org/10.1371/journal.pone.0105825

[75] Daniel Lakens. 2013. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology* 4 (2013), 12 pages.

[76] Daniël Lakens. 2014. Performing High-Powered Studies Efficiently with Sequential Analyses. *European Journal of Social Psychology* 44, 7 (2014), 701–710. https://doi.org/10.1002/ejsp.2023

[77] Daniël Lakens. 2022. Sample Size Justification. *Collabra: Psychology* 8, 1 (March 2022), 28. https://doi.org/10.1525/collabra.33267

[78] Daniël Lakens and Aaron R. Caldwell. 2021. Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science* 4, 1 (Jan. 2021), 2515245920951503. https://doi.org/10.1177/2515245920951503

[79] LangChain [n. d.]. LangGraph. https://www.langchain.com/langgraph.

[80] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* 36, 4 (Feb. 2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[81] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD Is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445488

[82] Mark W. Lipsey and David B. Wilson. 1993. The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis. *American Psychologist* 48, 12 (1993), 1181–1209.

[83] Mark W. Lipsey and David B. Wilson. 2001. *Practical Meta-Analysis*. Sage Publications, Inc, Thousand Oaks, CA, US. ix, 247 pages.

[84] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376533

[85] David Moher, Kenneth F. Schulz, and Douglas G. Altman. 2001. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials. *The Lancet* 357, 9263 (April 2001), 1191–1194. https://doi.org/10.1016/S0140-6736(00)04337-3

[86] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. https://doi.org/10.1145/3613904.3641936

[87] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[88] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. 2020. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376791

[89] Jakob Nielsen and Jonathan Levy. 1994. Measuring Usability: Preference vs. Performance. *Commun. ACM* 37, 4 (April 1994), 66–75. https://doi.org/10.1145/175276.175282

[90] Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods* 48, 4 (Dec. 2016), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

[91] Natalia Obukhova. 2021. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451520

[92] Stephen Olejnik and James Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4 (2003), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

[93] Open Science Collaboration. 2015. Estimating the Reproducibility of Psychological Science. *Science* 349, 6251 (2015), 9 pages. https://doi.org/10.1126/science.aac4716

[94] OpenAI. 2023. *GPT-4 Technical Report*. Technical Report. OpenAI.

[95] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3580766

[96] 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.

[97] Antti Oulasvirta and Kasper Hornbæk. 2016. HCI Research as Problem-Solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 4956–4967. https://doi.org/10.1145/2858036.2858283

[98] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3580841

[99] Shanta Pandey and Charlotte Lyn Bright. 2008. What Are Degrees of Freedom? *Social Work Research* 32, 2 (2008), 119–128. jstor:42659677

[100] Jaewoo Park, Eunji Lee, Yoonsik Kim, Isaac Kang, Hyung Il Koo, and Nam Ik Cho. 2020. Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *IEEE Access* 8 (2020), 174437–174448. https://doi.org/10.1109/ACCESS.2020.3025769

[101] Irene V. Pasquetto, Ashley E. Sands, Peter T. Darch, and Christine L. Borgman. 2016. Open Data in Scientific Settings: From Policy to Practice. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1585–1596. https://doi.org/10.1145/2858036.2858543

[102] Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445734

[103] Jose D. Perezgonzalez. 2015. Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing. *Frontiers in Psychology* 6 (March 2015), 11. https://doi.org/10.3389/fpsyg.2015.00223

[104] Chanda Phelan, Jessica Hullman, Matthew Kay, and Paul Resnick. 2019. Some Prior(s) Experience Necessary: Templates for Getting Started with Bayesian Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Chi '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300709

[105] Amy C Plint, David Moher, Andra Morrison, Kenneth Schulz, Douglas G Altman, Catherine Hill, and Isabelle Gaboury. 2006. Does the CONSORT Checklist Improve the Quality of Reports of Randomised Controlled Trials? A Systematic Review. *Medical Journal of Australia* 185, 5 (2006), 263–267. https://doi.org/10.5694/j.1326-5377.2006.tb00557.x

[106] Luke Plonsky and Frederick L Oswald. 2014. How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language learning* 64, 4 (2014), 878–912.

[107] Maciej P. Polak and Dane Morgan. 2024. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. *Nature Communications* 15, 1 (Feb. 2024), 1569. https://doi.org/10.1038/s41467-024-45914-8

[108] Leo Poom and Anders af Wåhlberg. 2022. Accuracy of Conversion Formula for Effect Sizes: A Monte Carlo Simulation. *Research Synthesis Methods* 13, 4 (2022), 508–519. https://doi.org/10.1002/jrsm.1560

[109] R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[110] Deepthi Raghunandan, Aayushi Roy, Shenzhi Shi, Niklas Elmqvist, and Leilani Battle. 2023. Code Code Evolution: Understanding How People Change Data Science Notebooks Over Time. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–12. https://doi.org/10.1145/3544548.3580997

[111] Jack Ratcliffe, Francesco Soave, Nick Bryan-Kinns, Laurissa Tokarchuk, and Ildar Farkhatdinov. 2021. Extended Reality (XR) Remote Research: A Survey of Drawbacks and Opportunities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 527, 13 pages. https://doi.org/10.1145/3411764.3445170

[112] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[113] John T. E. Richardson. 2011. Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research. *Educational Research Review* 6, 2 (Jan. 2011), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

[114] Katja Rogers and Katie Seaborn. 2023. The Systematic Review-lution: A Manifesto to Promote Rigour and Inclusivity in Research Synthesis. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–11. https://doi.org/10.1145/3544549.3582733

[115] Wade Rose and Inder Jit Singh Mann. 2011. The Variability of Pseudo R2s in Logistic Regression Models. *The IUP Journal of Computational Mathematics* IV, 1 (2011), 7–18.

[116] Robert Rosenthal. 1994. Parametric Measures of Effect Size. In *The Handbook of Research Synthesis*, Harris Cooper and Larry Hedges (Eds.). Russel Sage Foundation, New York, 231–244.

[117] Ralph L Rosnow and Robert Rosenthal. 1989. Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist* 44, 10 (1989), 1276–1284.

[118] Melvin T. Rupinski and William P. Dunlap. 1996. Approximating Pearson Product-Moment Correlations from Kendall's Tau and Spearman's Rho. *Educational and Psychological Measurement* 56, 3 (June 1996), 419–429. https://doi.org/10.1177/0013164496056003004

[119] David Saffo, Caglar Yildirim, Sara Di Bartolomeo, and Cody Dunne. 2020. Crowdsourcing Virtual Reality Experiments Using VRChat. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382829

[120] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580848

[121] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. 2023. Don't Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580791

[122] Thomas Schäfer and Marcus A. Schwarz. 2019. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology* 10, Article 813 (2019), 13 pages.

[123] Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. 2021. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science* 4, 2 (April 2021), 12 pages. https://doi.org/10.1177/25152459211007467

[124] Ruben Schlagowski, Dariia Nazarenko, Yekta Can, Kunal Gupta, Silvan Mertes, Mark Billinghurst, and Elisabeth André. 2023. Wish You Were Here: Mental and Physiological Effects of Remote Music Collaboration in Mixed Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581162

[125] Xenia Schmalz, José Biurrun Manresa, and Lei Zhang. 2023. What Is a Bayes Factor? *Psychological Methods* 28, 3 (June 2023), 705–718. https://doi.org/10.1037/met0000421

[126] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. https://doi.org/10.1145/3613904.3642459

[127] James P. Shaver. 1993. What Statistical Significance Testing Is, and What It Is Not. *The Journal of Experimental Education* 61, 4 (July 1993), 293–316. https://doi.org/10.1080/00220973.1993.10806592

[128] Thomas J Smith and Cornelius M McKenna. 2013. A Comparison of Logistic Regression Pseudo R2 Indices. *Multiple Linear Regression Viewpoints* 39, 2 (2013), 17–26.

[129] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[130] Shuyan Sun, Wei Pan, and Lihshing Leigh Wang. 2010. A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology. *Journal of Educational Psychology* 102, 4 (Nov. 2010), 989–1004. https://doi.org/10.1037/a0019507

[131] Mohammad Tahaei and Kami Vaniea. 2022. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3501957

[132] Ye Tao, Shuhong Wang, Junzhe Ji, Linlin Cai, Hongmei Xia, Zhiqi Wang, Jinghai He, Yitao Fan, Shengzhang Pan, Jinghua Xu, Cheng Yang, Lingyun Sun, and Guanyun Wang. 2023. 4Doodle: 4D Printing Artifacts Without 3D Printers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581321

[133] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. *Comput. Surveys* 55, 6 (Dec. 2022), 109:1–109:28. https://doi.org/10.1145/3530811

[134] Amalio Telenti, Michael Auli, Brian L. Hie, Cyrus Maher, Suchi Saria, and John P. A. Ioannidis. 2024. Large Language Models for Science and Medicine. *European Journal of Clinical Investigation* 54, 6 (2024), e14183. https://doi.org/10.1111/eci.14183 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/eci.14183

[135] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. https://doi.org/10.48550/arXiv.2309.10621 arXiv:2309.10621 [cs]

[136] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T. Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What If the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education. *Smart Learning Environments* 10, 1 (Feb. 2023), 15. https://doi.org/10.1186/s40561-023-00237-x

[137] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs]

[138] Transparent Statistics in Human-Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. https://transparentstats.github.io/guidelines. https://doi.org/10.5281/zenodo.1186169

[139] Transparent Statistics in Human-Computer Interaction Working Group. 2024. Transparent Statistics in HCI. https://transparentstatistics.org/.

[140] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley Pub. Co., Reading, Mass.

[141] Lucy Turner, Larissa Shamseer, Douglas G. Altman, Laura Weeks, Jodi Peters, Thilo Kober, Sofia Dias, Kenneth F. Schulz, Amy C. Plint, and David Moher. 2012. Consolidated Standards of Reporting Trials (CONSORT) and the Completeness of Reporting of Randomised Controlled Trials (RCTs) Published in Medical Journals. *Cochrane Database of Systematic Reviews* 2013, 11 (2012), 125 pages. https://doi.org/10.1002/14651858.mr000030.pub2

[142] Priyan Vaithilingam, Elena L. Glassman, Peter Groenwegen, Sumit Gulwani, Austin Z. Henley, Rohan Malpani, David Pugh, Arjun Radhakrishna, Gustavo Soares, Joey Wang, and Aaron Yim. 2023. Towards More Effective AI-assisted Programming: A Systematic Design Exploration to Improve Visual Studio IntelliCode's User Experience. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, Melbourne, Australia, 185–195. https://doi.org/10.1109/ICSE-SEIP58684.2023.00022

[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, USA, 11 pages.

[144] 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. https://doi.org/10.1145/3627508.3638298

[145] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2018. Special Interest Group on Transparent Statistics Guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3170427.3185374

[146] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems (NeurIPS'19, Vol. 32)*. Curran Associates, Inc., Vancouver, Canada, 15 pages.

[147] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. 2021. What Makes a Well-Documented Notebook? A Case Study of Data Scientists' Documentation Practices in Kaggle. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3451617

[148] 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. https://doi.org/10.1145/3613904.3641917

[149] Tony Wang, Haard K Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2023. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581372

[150] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond " $p < 0.05$". *The American Statistician* 73, sup1 (March 2019), 1–19. https://doi.org/10.1080/00031305.2019.1583913

[151] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 43 (2019), 1686. https://doi.org/10.21105/joss.01686

[152] Ziang Xiao, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2023. Inform the Uninformed: Improving Online Informed Consent Reading with an AI-Powered Chatbot. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581252

[153] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology* 55, 1 (2024), 90–112. https://doi.org/10.1111/bjet.13370

[154] Bereket A. Yilma and Luis A. Leiva. 2023. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581477

[155] Alyson L. Young, Wayne G. Lutters, Nicholas R. Magliocca, and Erle C. Ellis. 2013. Designing a System for Land Change Science Meta-Study. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. ACM, Paris France, 1473–1478. https://doi.org/10.1145/2468356.2468619

[156] Oleg Zendel, J. Shane Culpepper, Falk Scholer, and Paul Thomas. 2024. Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Sheffield United Kingdom, 340–345. https://doi.org/10.1145/3627508.3638322

[157] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, Toronto Ontario Canada, 110–120. https://doi.org/10.1145/3368555.3384448

[158] Guido Zuccon, Harrisen Scells, and Shengyao Zhuang. 2023. Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 283–289. https://doi.org/10.1145/3578337.3605121

# A Appendix Contents

The appendix contains the prompt used to extract statistics (see Section B) and Tables 6 to 14, which contain effect size guidelines for the remaining effect size measures used in the meta-study, $\eta^2$, Cohen's d, OR, CLES, Kendall's $\tau$, Spearman's $\rho$, Cramer's V, Cohen's f, and $f^2$. We provide these and the guidelines presented in the main paper in CSV-format for increased ease of processing in the supplemental material.

# B Prompt

Our final prompt consists of the following parts:

**System prompt incl. persona** You are a statistics expert and want to conduct a meta-analysis based on the results of scientific papers. To do this, you need to extract important statistics from the tests reported in the scientific paper. Carefully heed the user's instructions. Respond using Markdown.

**Task** You need to extract the number of participants in the study, i.e. the sample size. For tests in the null hypothesis statistical testing paradigma, you need the following information for each individual test:

- the name of the hypothesis test
- the p-value, which can take on numeric values between 0 and 1 and is often denoted as being smaller < or larger > than a given value
- the effect size, which is a different measure depending on the hypothesis test used, and consists of a measure, sometimes denoted by a Greek letter and the actual numeric effect size value
- confidence intervals around the effect size, which consist of two numeric values, one smaller than the effect size and one larger than the effect size. Confidence intervals can be different types, with 95% confidence intervals being the most common. They are frequently abbreviated CI
- the number of participants whose data was used in this test. This will often be same as the sample size in the study, but not always. E.g. in post-hoc tests for independent samples, only the participants in the conditions compared in the post-hoc tests are relevant for those post-hoc tests. Or when only participants who fulfill a certain condition are considered for an analysis.

Depending on the type of test, you need to extract additional information. Some tests will report multiple different effect sizes. You need to extract them all and if available, their confidence intervals. Some tests, like regression analyses, will report multiple p-values and corresponding effect sizes and confidence intervals, for each factor involved in the analysis. You need to collect the information for each factor separately, but it should still be identifiable to which test the factor belongs. Some tests, like analyses of variance (ANOVAs) will report results (i.e. p-values, corresponding effect sizes and confidence intervals) separately for one or more main effects and one or more interaction effects. You need to collect the information for each effect separately, but it should still be identifiable to which test the factor belongs.

A combination of these is also possible, so that a test can have a p-value, effect sizes and confidence intervals associated with the test in general, also called omnibus test, and additional p-values, effect sizes and confidence intervals associated with each individual factor and effect in the test.

Any of the numeric values, such as p-values, effect sizes or confidence intervals can be missing. Sometimes these values are not stated explicitly for each test, e.g. tests with the same result may be summarized e.g. as: "All other comparisons were not significant". In that case, it is necessary to identify

how many tests were conducted in total, e.g. by comparing how many conditions there are in total and how many are not yet accounted for by the reporting. Then list each non-significant test separately. If the numeric values cannot be extrapolated from the context, they should be considered missing and you need to state explicitly that the value is missing.

Other values (like names of hypothesis tests) may also be missing, in which case determine the value from the context as best as you can. Test statistics or descriptions in the method section, regarding the used tests in the data analysis or the study set-up (e.g. to determine whether repeated measures analyses were likely used, or which types of variables were measured) can be helpful.

Some values are only reported in tables. If the table's title or a reference in the text suggests relevant information in the table request it via the tool!

If the paper reports tests in a different statistical paradigm, e.g. using Bayes statistics, you also need information for each individual test:

- the name of the hypothesis test
- all other statistics associated with this hypothesis test, including names and values of the statistics

**Paper information** The paper you have to analyze has the title [title]. Its abstract is the following:

"' [abstract] "'

It has the following sections:

"' [sections] "'

And the following tables:

"' [tables] "'

Analyze the paper using your tools and make sure to report every test with its parameters. Remember to also report effect sizes that are not explicitly stated including but not limited to:

- correlation coefficients.
- Odds Ratio (OR)
- Risk Ratio (RR)
- Related Measures

After extracting the information for a test (test name, effect size, amount participants, ...), **use the 'report' tool immediately** to submit the test. Once the report is submitted, confirm with 'Report submitted' before proceeding to the next test. Do not move forward until you have completed the report.

After reviewing the sections, have a look at the relevant tables.

## C Effect Size Guidelines

We provide the effect size guidelines for the remaining effect sizes measures used in the meta-study: $\eta^2$, Cohen's d, OR,CLES, Kendall's $\tau$, Spearman's $\rho$, Cramer's V, Cohen's f and $f^2$.

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.02 | 0.07 | 0.37 | 0.01 | 0.10 | 0.52 |
| applied computing | 0.02 | 0.11 | 0.46 | 0.02 | 0.19 | 0.56 |
| collaborative and social computing | 0.00 | 0.07 | 0.37 | 0.00 | 0.07 | 0.38 |
| computing methodologies | 0.00 | 0.04 | 0.49 | 0.00 | 0.10 | 0.93 |
| empirical studies in hci | 0.02 | 0.07 | 0.34 | 0.01 | 0.08 | 0.36 |
| hci theory, concepts and models | 0.02 | 0.10 | 0.52 | 0.01 | 0.36 | 0.97 |
| human computer interaction (hci) | 0.01 | 0.09 | 0.42 | 0.03 | 0.11 | 0.41 |
| human-centered computing | 0.02 | 0.07 | 0.37 | 0.01 | 0.10 | 0.55 |
| interaction design | 0.02 | 0.06 | 0.08 | 0.01 | 0.02 | 0.98 |
| mixed / augmented reality | 0.08 | 0.15 | 0.37 | 0.06 | 0.23 | 0.60 |
| security and privacy | 0.02 | 0.05 | 0.16 | 0.00 | 0.03 | 0.20 |
| social and professional topics | 0.02 | 0.04 | 0.42 | 0.04 | 0.21 | 0.64 |
| virtual reality | 0.07 | 0.10 | 0.25 | 0.07 | 0.28 | 0.53 |

**Table 6: Guidelines for separate research areas, using $\eta^2$, $R^2$, $\omega^2$, $\eta_p^2$, $\eta_G^2$ or related effect size measures**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.22 | 0.44 | 1.09 | 0.19 | 0.52 | 1.38 |
| applied computing | 0.20 | 0.54 | 1.26 | 0.25 | 0.74 | 1.46 |
| collaborative and social computing | 0.11 | 0.44 | 1.09 | 0.08 | 0.42 | 1.12 |
| computing methodologies | 0.11 | 0.32 | 1.32 | 0.07 | 0.52 | 2.33 |
| empirical studies in hci | 0.22 | 0.43 | 1.04 | 0.17 | 0.46 | 1.07 |
| hci theory, concepts and models | 0.22 | 0.51 | 1.38 | 0.18 | 1.07 | 2.46 |
| human computer interaction (hci) | 0.19 | 0.49 | 1.18 | 0.27 | 0.54 | 1.17 |
| human-centered computing | 0.22 | 0.44 | 1.09 | 0.19 | 0.52 | 1.44 |
| interaction design | 0.23 | 0.41 | 0.46 | 0.13 | 0.20 | 2.50 |
| mixed / augmented reality | 0.46 | 0.64 | 1.08 | 0.40 | 0.81 | 1.55 |
| security and privacy | 0.23 | 0.36 | 0.65 | 0.09 | 0.27 | 0.75 |
| social and professional topics | 0.21 | 0.34 | 1.18 | 0.34 | 0.77 | 1.61 |
| virtual reality | 0.42 | 0.52 | 0.86 | 0.42 | 0.92 | 1.40 |

**Table 7: Guidelines for separate research areas, using Cohen's d**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 1.48 | 2.13 | 5.90 | 1.39 | 2.44 | 8.72 |
| applied computing | 1.42 | 2.50 | 7.46 | 1.54 | 3.47 | 9.63 |
| collaborative and social computing | 1.21 | 2.12 | 5.93 | 1.16 | 2.06 | 6.12 |
| computing methodologies | 1.21 | 1.74 | 8.10 | 1.13 | 2.44 | 23.93 |
| empirical studies in hci | 1.48 | 2.11 | 5.46 | 1.35 | 2.19 | 5.71 |
| hci theory, concepts and models | 1.46 | 2.38 | 8.68 | 1.37 | 5.71 | 26.48 |
| human computer interaction (hci) | 1.39 | 2.30 | 6.67 | 1.59 | 2.52 | 6.57 |
| human-centered computing | 1.46 | 2.13 | 5.93 | 1.39 | 2.44 | 9.39 |
| interaction design | 1.50 | 2.02 | 2.21 | 1.25 | 1.43 | 27.38 |
| mixed / augmented reality | 2.22 | 2.99 | 5.84 | 1.99 | 3.89 | 10.76 |
| security and privacy | 1.50 | 1.86 | 3.03 | 1.17 | 1.59 | 3.53 |
| social and professional topics | 1.43 | 1.79 | 6.67 | 1.78 | 3.67 | 11.63 |
| virtual reality | 2.07 | 2.46 | 4.17 | 2.06 | 4.55 | 8.94 |

**Table 8: Guidelines for separate research areas, using OR**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.54 | 0.59 | 0.71 | 0.54 | 0.60 | 0.76 |
| applied computing | 0.54 | 0.61 | 0.74 | 0.55 | 0.64 | 0.77 |
| collaborative and social computing | 0.52 | 0.59 | 0.71 | 0.52 | 0.58 | 0.71 |
| computing methodologies | 0.52 | 0.56 | 0.75 | 0.51 | 0.60 | 0.92 |
| empirical studies in hci | 0.54 | 0.59 | 0.70 | 0.54 | 0.59 | 0.70 |
| hci theory, concepts and models | 0.54 | 0.60 | 0.76 | 0.54 | 0.70 | 0.94 |
| human computer interaction (hci) | 0.54 | 0.60 | 0.72 | 0.55 | 0.61 | 0.72 |
| human-centered computing | 0.54 | 0.59 | 0.71 | 0.54 | 0.60 | 0.77 |
| interaction design | 0.55 | 0.58 | 0.59 | 0.53 | 0.54 | 0.96 |
| mixed / augmented reality | 0.59 | 0.63 | 0.71 | 0.58 | 0.66 | 0.78 |
| security and privacy | 0.55 | 0.57 | 0.63 | 0.52 | 0.55 | 0.65 |
| social and professional topics | 0.54 | 0.57 | 0.72 | 0.57 | 0.65 | 0.79 |
| virtual reality | 0.58 | 0.60 | 0.67 | 0.58 | 0.68 | 0.76 |

**Table 9: Guidelines for separate research areas, using CLES**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.09 | 0.17 | 0.41 | 0.08 | 0.20 | 0.51 |
| applied computing | 0.08 | 0.21 | 0.47 | 0.10 | 0.29 | 0.53 |
| collaborative and social computing | 0.04 | 0.17 | 0.41 | 0.03 | 0.17 | 0.42 |
| computing methodologies | 0.04 | 0.13 | 0.49 | 0.03 | 0.20 | 0.81 |
| empirical studies in hci | 0.09 | 0.17 | 0.39 | 0.07 | 0.18 | 0.40 |
| hci theory, concepts and models | 0.09 | 0.20 | 0.50 | 0.07 | 0.40 | 0.86 |
| human computer interaction (hci) | 0.08 | 0.19 | 0.44 | 0.11 | 0.21 | 0.44 |
| human-centered computing | 0.09 | 0.17 | 0.41 | 0.08 | 0.20 | 0.52 |
| interaction design | 0.09 | 0.16 | 0.18 | 0.05 | 0.08 | 0.88 |
| mixed / augmented reality | 0.18 | 0.25 | 0.41 | 0.16 | 0.31 | 0.56 |
| security and privacy | 0.09 | 0.14 | 0.26 | 0.04 | 0.11 | 0.29 |
| social and professional topics | 0.08 | 0.14 | 0.44 | 0.13 | 0.30 | 0.58 |
| virtual reality | 0.17 | 0.21 | 0.33 | 0.17 | 0.35 | 0.51 |

**Table 10: Guidelines for separate research areas, using Kendall's $\tau$**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.13 | 0.26 | 0.58 | 0.11 | 0.30 | 0.69 |
| applied computing | 0.12 | 0.31 | 0.65 | 0.15 | 0.41 | 0.72 |
| collaborative and social computing | 0.06 | 0.26 | 0.58 | 0.05 | 0.24 | 0.59 |
| computing methodologies | 0.06 | 0.19 | 0.67 | 0.04 | 0.30 | 0.94 |
| empirical studies in hci | 0.13 | 0.25 | 0.56 | 0.10 | 0.27 | 0.57 |
| hci theory, concepts and models | 0.13 | 0.29 | 0.69 | 0.11 | 0.57 | 0.96 |
| human computer interaction (hci) | 0.11 | 0.28 | 0.61 | 0.16 | 0.31 | 0.61 |
| human-centered computing | 0.13 | 0.26 | 0.58 | 0.11 | 0.30 | 0.71 |
| interaction design | 0.14 | 0.24 | 0.27 | 0.08 | 0.12 | 0.97 |
| mixed / augmented reality | 0.27 | 0.37 | 0.57 | 0.23 | 0.45 | 0.75 |
| security and privacy | 0.14 | 0.21 | 0.37 | 0.05 | 0.16 | 0.42 |
| social and professional topics | 0.12 | 0.20 | 0.61 | 0.20 | 0.43 | 0.77 |
| virtual reality | 0.25 | 0.30 | 0.47 | 0.25 | 0.50 | 0.70 |

**Table 11: Guidelines for separate research areas, using Spearman's $\rho$**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.10 | 0.19 | 0.42 | 0.08 | 0.22 | 0.50 |
| applied computing | 0.09 | 0.23 | 0.47 | 0.11 | 0.31 | 0.52 |
| collaborative and social computing | 0.05 | 0.19 | 0.42 | 0.04 | 0.18 | 0.43 |
| computing methodologies | 0.05 | 0.14 | 0.49 | 0.03 | 0.22 | 0.67 |
| empirical studies in hci | 0.10 | 0.19 | 0.41 | 0.08 | 0.20 | 0.42 |
| hci theory, concepts and models | 0.10 | 0.22 | 0.50 | 0.08 | 0.42 | 0.68 |
| human computer interaction (hci) | 0.08 | 0.21 | 0.45 | 0.12 | 0.23 | 0.44 |
| human-centered computing | 0.10 | 0.19 | 0.42 | 0.08 | 0.22 | 0.52 |
| interaction design | 0.10 | 0.18 | 0.20 | 0.06 | 0.09 | 0.69 |
| mixed / augmented reality | 0.20 | 0.27 | 0.42 | 0.17 | 0.33 | 0.54 |
| security and privacy | 0.10 | 0.16 | 0.27 | 0.04 | 0.12 | 0.31 |
| social and professional topics | 0.09 | 0.15 | 0.45 | 0.15 | 0.32 | 0.55 |
| virtual reality | 0.18 | 0.22 | 0.35 | 0.18 | 0.37 | 0.51 |

**Table 12: Guidelines for separate research areas, using Cramer's V**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.11 | 0.22 | 0.55 | 0.10 | 0.26 | 0.69 |
| applied computing | 0.10 | 0.27 | 0.63 | 0.12 | 0.37 | 0.73 |
| collaborative and social computing | 0.06 | 0.22 | 0.55 | 0.04 | 0.21 | 0.56 |
| computing methodologies | 0.06 | 0.16 | 0.66 | 0.04 | 0.26 | 1.17 |
| empirical studies in hci | 0.11 | 0.22 | 0.52 | 0.09 | 0.23 | 0.53 |
| hci theory, concepts and models | 0.11 | 0.25 | 0.69 | 0.09 | 0.53 | 1.23 |
| human computer interaction (hci) | 0.10 | 0.24 | 0.59 | 0.13 | 0.27 | 0.58 |
| human-centered computing | 0.11 | 0.22 | 0.55 | 0.10 | 0.26 | 0.72 |
| interaction design | 0.12 | 0.20 | 0.23 | 0.06 | 0.10 | 1.25 |
| mixed / augmented reality | 0.23 | 0.32 | 0.54 | 0.20 | 0.41 | 0.78 |
| security and privacy | 0.12 | 0.18 | 0.33 | 0.04 | 0.13 | 0.38 |
| social and professional topics | 0.10 | 0.17 | 0.59 | 0.17 | 0.39 | 0.81 |
| virtual reality | 0.21 | 0.26 | 0.43 | 0.21 | 0.46 | 0.70 |

**Table 13: Guidelines for separate research areas, using Cohen's f**

| research area | between groups | | | within groups | | |
|---|---|---|---|---|---|---|
| | small | medium | large | small | medium | large |
| all | 0.02 | 0.08 | 0.59 | 0.01 | 0.11 | 1.08 |
| applied computing | 0.02 | 0.12 | 0.85 | 0.02 | 0.24 | 1.27 |
| collaborative and social computing | 0.00 | 0.08 | 0.59 | 0.00 | 0.07 | 0.62 |
| computing methodologies | 0.00 | 0.04 | 0.96 | 0.00 | 0.11 | 13.29 |
| empirical studies in hci | 0.02 | 0.08 | 0.52 | 0.01 | 0.09 | 0.56 |
| hci theory, concepts and models | 0.02 | 0.11 | 1.08 | 0.01 | 0.56 | 32.33 |
| human computer interaction (hci) | 0.01 | 0.10 | 0.71 | 0.03 | 0.12 | 0.69 |
| human-centered computing | 0.02 | 0.08 | 0.59 | 0.01 | 0.11 | 1.22 |
| interaction design | 0.02 | 0.07 | 0.09 | 0.01 | 0.02 | 57.82 |
| mixed / augmented reality | 0.09 | 0.18 | 0.58 | 0.07 | 0.30 | 1.53 |
| security and privacy | 0.02 | 0.05 | 0.18 | 0.00 | 0.03 | 0.25 |
| social and professional topics | 0.02 | 0.05 | 0.71 | 0.05 | 0.27 | 1.75 |
| virtual reality | 0.07 | 0.12 | 0.33 | 0.07 | 0.39 | 1.13 |

**Table 14: Guidelines for separate research areas, using $f^2$**