Statslator: Interactive Translation of NHST and Estimation Statistics Reporting Styles in Scientific Documents

Damien Masson, Sylvain Malacria, Géry Casiez, Daniel Vogel

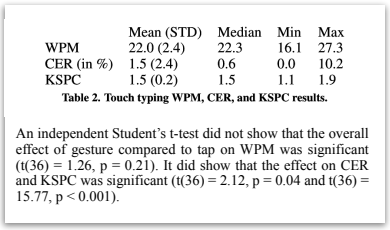
Statslator: Interactive Translation of NHST and Estimation Statistics Reporting Styles in Scientific Documents

Damien Masson
Cheriton School of Computer Science, University of Waterloo
Waterloo, Canada
dmasson@uwaterloo.ca

Sylvain Malacria*
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL
Lille, France
sylvain.malacria@inria.fr

Géry Casiez†‡
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
Lille, France
gery.casiez@univ-lille.fr

Daniel Vogel
Cheriton School of Computer Science, University of Waterloo
Waterloo, Canada
dvogel@uwaterloo.ca



	Mean (STD)	Median	Min	Max
WPM	22.0 (2.4)	22.3	16.1	27.3
CER (in %)	1.5 (2.4)	0.6	0.0	10.2
KSPC	1.5 (0.2)	1.5	1.1	1.9

Table 2. Touch typing WPM, CER, and KSPC results.

An independent Student's t-test did not show that the overall effect of gesture compared to tap on WPM was significant ($t(36) = 1.26$, $p = 0.21$). It did show that the effect on CER and KSPC was significant ($t(36) = 2.12$, $p = 0.04$ and $t(36) = 15.77$, $p < 0.001$).

(a) NHST Statistical Report

Measure	*Gesture*	*Tap*
N	18	18
Mean	24	22
SD	5.8	2.4
SE	1.37	0.57
95% MoE	2.88	1.19

Measure	*Gesture - Tap*
Paired?	0
df	34
Simple ES	2
t	1.28
p-value	0.21
95% MoE	3.18
Cohen's d	0.45
CLES	0.62
rpb	0.22
OR	2.26
S-value	2.25

(b) Converted Report

Figure: Dot plot of the mean difference. Error bars represent the 95% CI.

(c) Graphical Report

Figure 1: Statslator takes existing statistical reports (a) using NHST or estimation; (b) calculates all possible statistical values using accurate conversion equations; (c) shows the report using graphical and interactive figures configurable by readers.

ABSTRACT

Inferential statistics are typically reported using p-values (NHST) or confidence intervals on effect sizes (estimation). This is done using a range of styles, but some readers have preferences about how statistics should be presented and others have limited familiarity with alternatives. We propose a system to interactively translate statistical reporting styles in existing documents, allowing readers to switch between interval estimates, p-values, and standardized effect sizes, all using textual and graphical reports that are dynamic and user customizable. Forty years of CHI papers are examined. Using only the information reported in scientific documents, equations are derived and validated on simulated datasets to show that conversions between p-values and confidence intervals are accurate. The system helps readers interpret statistics in a familiar style, compare reports that use different styles, and even validate the correctness of reports. Code and data: https://osf.io/x4ue7

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

statistics, interactive system, reading interface, estimation, nhst, transparent statistics, explorable explanation

ACM Reference Format:
Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2023. Statslator: Interactive Translation of NHST and Estimation Statistics Reporting Styles in Scientific Documents. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3586183.3606762

1 INTRODUCTION

When studying a population, a common approach is to collect data from a random sample of that population (e.g., participants) and use inferential statistics to generalize the findings. A detailed report of this analysis in scientific documents enables readers to evaluate the strength of the findings. However, the presentation of these reports as imposed by authors might be hard to understand [20, 65], incomplete [94], even misleading [7, 41, 68].

Within the HCI community and other empirical fields, statistical reports commonly include *p-values* obtained from *null hypothesis significance testing* (NHST). However, p-values and their associated

*Also with University of Waterloo.
†Also with Institut Universitaire de France.
‡Also with University of Waterloo.

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA

reporting language might mislead even trained scientists [68, 74], and reporting p-values alone is insufficient to draw meaningful and nuanced conclusions, capture uncertainty of results, and answer quantitative questions about the effect [21, 27, 41, 67]. For these reasons, some recommend reporting additional information such as effect sizes and interval estimates [5, 94], while others advocate for avoiding p-values and NHST altogether [15, 20]. This debate is particularly visible in HCI where all positions cohabit. P-values are prevalent in publications [8, 97] and more robust NHST approaches are being developed [35, 99]. Yet, workshops and articles are also promoting estimation [15, 27, 43] and bayesian approaches [61, 81].

While much of the ongoing debate is about *what* to report, it also extends to *how* inferential statistics should be reported. For example, what symbols to use [5, sec 6.44], what words to communicate findings [9, 27], and what numeric precision to use [10]. How authors report statistics continuously evolves as the field matures: for example, the practise of reporting p-values using inequalities has noticeably decreased since 2010 [8]. New reporting styles like graphical presentations of effect sizes can effectively convey the information [15, 43, 48], but such figures are seldom found in documents because they can be difficult to create [17, 60] and take additional space [2, 43]. Even graphical representations can be misunderstood, for example, error bars may depict standard deviations, standard errors, and confidence intervals [22]. At best, the meaning of error bars is clear and a reader can recover a quantity of interest by "eyeballing it" [21, ch 5]. At worst, a reader can form the wrong conclusion about the depicted results [65].

It is virtually impossible to please all readers with one style of report for inferential analysis. For this reason, some recommend including multiple reporting styles [5, sec 3.7], or even multiple alternative analysis ("multiverse analysis") [45, 92] which can be made interactive to give readers control over the presentation [28]. However, they are difficult to create [44, 70, 88] and often incompatible with publishing workflows that impose page limits and digital formats with no support for interactivity [72]. Even in the unlikely case that all authors adopt reporting styles akin to interactive multiverse analyses, it is unclear how previously published documents could be supported, especially when their raw data is unavailable.

We argue for reader-centred statistics: readers should have the final word on statistical inference presentation because what matters is correct interpretation, be it through estimation or NHST[1], textual or graphical content, or static or interactive documents. Moreover, this should be possible with old and new documents without substantial efforts. Our key insight is that most scientific articles report enough information about at least one type of inferential statistical analysis which can be translated into a different reporting style. For example, a confidence interval (CI) can be calculated given a p-value and means. The difficulty lies in three aspects: (1) obtaining the statistical information—we analyzed statistical reports at CHI and propose a semi-automatic pipeline to extract common statistical reports; (2) converting the information into a target statistic—we demonstrate how to do these conversions and thoroughly evaluate their accuracy given common reporting practices; and (3) presenting the converted information back to readers in their preferred style—we design a document reader that allows readers to customize textual, graphical, and interactive statistical reports. We implement our solutions to these aspects into Statslator, a tool for readers to retro-actively encode best practises in terms of statistical reporting with no author involvement. This supports documents that were already disseminated and bypasses limitations due to outdated publishing practises that discourage the use of figures because of page limits and prevent interactive documents due to reliance on the PDF format or inflexible publishing workflows.

[1]Some argue p-values cannot be properly interpreted even when their pitfalls are understood [15]. We believe readers should still decide if they wish to use p-values.

2 BACKGROUND AND RELATED WORK

We first review the challenges associated with different ways of reporting statistics and then detail recommendations made by the community to tackle these challenges. We then detail how previous systems allowed readers to personalize documents without involving authors.

2.1 Challenges with Statistical Reporting

Prevalent practices in scientific documents can make statistical reports hard to understand.

First, *there is much confusion about the meaning of statistical values.* For example, the potential for misinterpretating p-values [21, 41]: a p-value might lead a reader to believe the result is more certain than it is (dichotomous thinking); that it conveys effect size (ambiguous use of "significant"); and that there is no difference when p>.05 or that p is the probability that the null hypothesis is true (inverse probability fallacy). But misinterpretations are not limited to p-values. Any statistical report not fully understood by readers might be misinterpreted, including standard errors, confidence intervals [7], and effect sizes [62]. Thus, it is important to give readers access to statistics they are familiar with, or default to ones that are easily understood.

Second, *there are many equivalent ways to report the same statistical values, all requiring a different interpretation.* Consider two groups $M_1 = 5, SD = 1$ and $M_2 = 7, SD_2 = 3$. Now consider the multitude of valid ways to convey the effect size (calculated value in parenthesis): the mean difference (2), Cohen's d (.89), Glass' δ (.67), the rank-biserial correlation (.41), odds ratio (5.1), η^2 (.17), Cohen's f (.45), and the common language effect size (.74). Worse, there is a lack of consensus on how to calculate some specific effect sizes such as Cohen's d and all variations are found in documents, with no indication of what formula was used (like we just did) [66]. Similarly, a confidence interval can be reported at various confidence level, and for different estimates [21, p 118]. These inconsistencies defeat most of the purpose of standardized effect sizes to convey an effect size that is comparable across studies.

Third, *there are many ways to report and represent statistical values.* In textual reports, values are often introduced with letters and Greek symbols [5, sec 6.44]. Besides being confusing, these symbols are inconsistent across documents and sometimes collide. For example, d can refer to any of the four ways of calculating Cohen's d [66]. Similarly, r is the symbol for both the rank-biseral correlation and the Pearson correlation coefficient (which are equivalent only in specific situations). The reported values themselves can be rounded with arbitrary precision, or they can be given as inequalities, often the case for p-values [8, 10].

Similar inconsistencies are in graphical reports of statistics. There is a wide diversity of chart types, and much has been written about how charts can deceive or confuse an inattentive reader [13, 79]. Additionally, standard graphical marks may not have consistent meanings. For example, the visual style of an "error bar" is easily recognizable, but whether they represent standard error, standard deviation or a confidence interval may be misinterpreted or unknown [7, 22, 51, 65]. Even the simple bars or points could be representing separate means or mean differences [20]. With adequate detail in a caption, these kinds of issues can be mitigated and some recommend graphical reports instead of text [27, 38, 43, 48]. But, authors are reluctant to include many graphs since they can be difficult and time consuming to create and take up additional space [2, 17, 60], which is unwanted for publications that limit page size or for peer review guidelines that equate contribution size to paper length.

Statistical reports are already difficult to read and understand for readers, and the diversity and inconsistency in how statistics are reported only makes it more challenging. Our approach enables a customizable presentation of statistical results controlled by readers, with embedded context, consistent calculations, and connections across reporting styles.

2.2 Recommendations to Report Statistics

The primary response to challenges with statistical reporting is to encourage authors to present results using a consistent, detailed, and clear style. For example, the American Psychological Association (APA) states that *"complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals"* [5, sec 3.7]. To make these reports consistent, APA also recommends specific phrasings and symbols. Within the HCI community, similar recommendations have been proposed [27, 43], and tools have been created to assist authors in the process [57, 98]. But these guidelines have yet to become standard practice [39, 40]: as of 2018, about 15% of CHI papers included confidence interval whereas 50% reported p-values [8].

In parallel, these recommendations keep evolving. For example, the APA started recommending the use of CIs in its fifth edition following a push from the community [46, 64]. Alternative presentations of statistics have been proposed and are often supported by experimental evidence. For example, figures showing effect size and confidence intervals can help readers [27, 43, 48]. Hypothetical outcome plots [56] (HOPs) rely on animations and help convey uncertainty [59]. Multiverse analysis reports [45, 92] that might be explorable [28], can highlight how fragile or strong the results are. Analogies and some more natural effect sizes such as the common language effect size are often better understood [62].

While recommendations exist, the bottleneck seems to be in their implementation, either because of slow adoption, difficulty of creation, or publishing format limitations. And of course, even if all these issues were to be solved, the problem would remain for existing documents.

2.3 Personalized Reading Experience

When authors' adoption of new guidelines is slow or unlikely to happen, a possible solution is to offer tools for readers. Several systems have been proposed that take as input a document and augment it in various ways.

For example, reported measurements such as distances might be difficult to interpret if readers cannot relate to them. Thus, systems have been proposed to automatically generate analogies and relatable explanations and visualizations of the measurements reported in a document [55, 63]. Similarly, there is a wealth of research on how to generate visualizations to accompany documents, either to give more contexts while reading by leveraging external databases of relevant information [37, 54], to generate visualization in-context for data that is already in the document but scattered in textual tables [6], or simply to give readers a way to annotate documents using charts so that they can make sense of numbers in-text [71].

These approaches have the advantage of being immediately applicable and to all documents, including those already disseminated. In this work, we adapt this approach to the context of statistical reports, and adjust the generated presentations to fit the numerous recommendations made by the community.

3 WHAT IS REPORTED AT CHI?

We analyzed the proceedings of ACM CHI conferences to better understand what inferential statistics are reported and how they are presented. The goal is to gauge the feasibility of generating different statistical representations from the data reported in text. We examine CHI papers because HCI is a multidisciplinary field and CHI is very large and diverse. Statistical practises likely vary among CHI authors depending on their field, background, and exposure to inferential statistics.

Our analysis focuses on text, not values included within figures or tables, and our approach does not consider complex sentence structures. Therefore, our results should be viewed as lower bounds rather than absolute proportions. The code and data of our analysis is accessible online: https://osf.io/x4ue7/

3.1 Corpus of Papers and Analysis

We scraped 9,611 PDFs from 1982 to 2022[2]. For each paper, text was extracted using a Python port[3] of MuPDF[4] and a set of case-insensitive regular expressions as identified statistical reports:

- One way to identify reported values is to look for numbers preceded by the relationship symbol '=', '<', '>', '≤', '≥', '<<', or '>>'. When found, the number and preceding word or symbol were extracted.
- Another way to identify values reported in text without a relationship symbol is to look for names of common descriptive and inferential statistics such as 'mean', 'median', 'standard deviation' and variations such as 'M', 'Mdn', and 'μ'. We only extracted those followed by the verb 'be' in any form such as 'is', 'was', and 'were', and then followed by a number.
- To count the mention of confidence intervals, we looked for the terms '%CI', '% CI', and 'confidence intervals'.
- To detect statistical tests, we checked for the name of a specific test among a list of 15 popular ones, including 't-test', 'ANOVA' and possible variations such as 'ANCOVA' and 'MANOVA'.

[2]Note CHI was not held in 1984
[3]https://github.com/pymupdf/PyMuPDF
[4]https://mupdf.com/

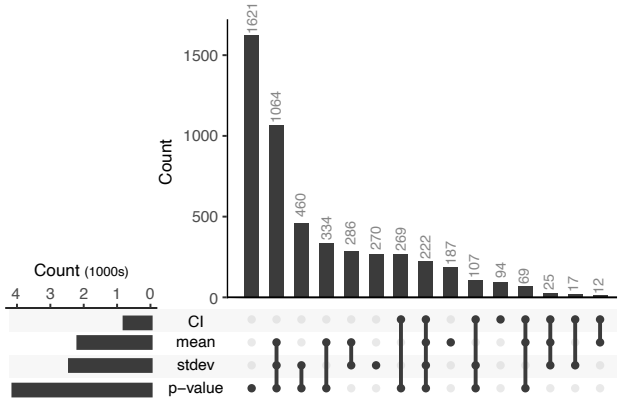


Figure 2: Upset plot showing the number of papers that report different sets of CIs, means, standard deviations, and p-values. Lines connecting dots across rows indicate what values are in the intersecting set.

The results were harmonized by reviewing the top candidates and grouping the ones referring to identical values. For example, 'mean', 'average', 'M', and 'μ' were all grouped as 'mean'. Statistical tests known under different names were also grouped, such as 'Wilcoxon Rank Sum Test' and 'Mann-Whitney U Test'.

3.2 Results

As a sanity check, we compared our results to Besançon and Dragicevic [8] who also used regular expressions to examine CHI proceedings although their analysis was limited to p-values and CIs from 2010 and 2018. We found 5.9% to 14.2% of CI for 2010-2018, they report "from 6% to 15%". We found 48.4% of p-values, they report "around 50%".

3.2.1 Reported Values. 6,266 papers (65%) mentioned at least one value. P-values (66%) were most common, followed by standard deviations (39%), means (35%) , F-values (28%), t-scores (19%), and confidence intervals (13%). Standard errors were found in only 2% of papers. Standardized effect sizes were also seldom found, such as Pearson's correlation coefficient (11%) and Cohen's d (5%).

Figure 2 shows the number of times a CI, mean, standard deviation, or p-value, or combination of these values, are reported in a paper. Of particular interest are papers reporting both p-values and means; means and standard deviations; and CI and means. For most of these 2,012 papers reporting p-values and means, we will show translating to a different reporting style is possible (section 4). Although papers reporting only p-values (1,612) cannot be translated to CIs, standardized effect sizes can still be calculated to complete the report (see section 4.3.1). For the remaining 2,642 papers, additional values will be needed. Overall, a large proportion of CHI papers have useful statistical reports, and recall that our conservative analysis likely underestimate actual occurrences and it is possible that values like means can be estimated from figures [58, 73]

3.2.2 Number of Decimals. Consistent with APA recommendations [5, sec 6.36], standard deviations, means, F-values, t-scores, and CIs were reported with a median of two decimals. The exception was for p-values, which were reported with a median of three decimals. However, 17% of these p-values were reported as inequalities with values '0.05' or '0.01'.

3.2.3 Statistical Tests. 4,800 papers (50%) mentioned at least one statistical test. Among these papers, t-tests were mentioned in most (61%), followed by ANOVA (49%), Wilcoxon signed-rank test (10%), Mann-Whitney U test (9%), and Chi-squared test (9%). Other tests, such as Friedman, were found in less than 5% of these papers.

4 CONVERTING STATISTICAL REPORTS

Based on the information reported in papers, we present a set of equations to perform bidirectional conversions between NHST-based reports and estimation-based reports. To our knowledge, these derivations were not or superficially covered, especially in the context of practices common in HCI studies. For example, some readers might have been taught the "conversion rule" that 95% CI is about twice as large as the standard error; the "overlap rule" that if two independent 95% CIs on the separate mean just touch, p is about 0.01, and no overlap means p<0.01; and the "difference rule" that if p>.05 then the 95% CI on the effect size will extend slightly past 0 [21, p 183]. What is sometimes omitted, however, is how these methods were derived, how precise conversions can be obtained, and when they do not work. For example, the overlap rule for a CI fails when the study follows a within-subject design [21, p 200], and the other two assume large sample sizes. Similarly, a widely cited article in the medical science community described how to convert p-values into CIs [3], but it assumes more than 60 participants: if applied to smaller studies typical in HCI [12], the converted estimations would be overoptimistic.

For all conversions, we presume the size of the groups N_1 and N_2 and the study design are known (i.e., *within* or *between-subject*) since this information would be reported in any rigorous scientific report. By extension, the degrees of freedom df for a t-test can be calculated if not already reported. For within designs $df = N - 1$, otherwise $df = N_1 + N_2 - 2$. Unless specified otherwise, equations are valid for all variations of t-test, within and between subject, groups of various sizes, and with equal or unequal variances.

4.1 Converting to Confidence Intervals

When comparing two groups, the CI of interest is the CI for the effect size, where the effect size is usually the difference between the two group means. This CI can replace a p-value as it conveys the estimate of effect size and the uncertainty around it.

The CI is calculated as $[\Delta M - MoE, \Delta M + MoE]$ where $\Delta M = M_2 - M_1$ is the difference of the group means, or "unstandardized" effect size, and MoE is the margin of error (corresponding to half the CI). Below, we show different ways to calculate MoE needed to obtain the CI.

4.1.1 From means and t-score. Cumming [21, p 163] explains that the calculation of the MoE depends on the t component $C_t = t_\alpha(df)$, the variability component C_v, and the sample size component C_s. Adopting this terminology, the calculation of a t-score can be expressed as follows.

$$tscore = \frac{M_2 - M_1}{C_v \times C_s} \quad (1)$$

Note this formulation is an abstraction since the calculation of C_v and C_s depends on the choice of t-test and the study design. But by referring to these three components, our equations are compatible with all standard t-tests. As such, the equation above can be rearranged to recover $C_v \times C_s$ given M_1 and M_2 and a t-score. The remaining C_t term is calculated using the t-distribution for a given degree of freedom (noted $t(df)$). As a result, given a t-score $tscore$, the means of both groups M_1 and M_2, the MoE at confidence level α is calculated as follows

$$MoE_\alpha = t_\alpha(df) \times \frac{M_2 - M_1}{tscore} \tag{2}$$

A simplification that is often made is to use the normal distribution instead of the t-distribution; a large sample size results in large degrees of freedom, in which case the t-distribution approximates a normal distribution, so knowing df can be relaxed. For example, for a large df and a 95% confidence level, $t_{.95} \approx z_{.95} = 1.96$ which simplifies to the equation presented by Altman and Bland [3]. However, in many HCI studies, the sample size is small, frequently around 12 participants [12]. In these cases, this assumption leads to narrower "overoptimistic" CIs and should be avoided.

4.1.2 From means and p-value. The p-value of a t-test is measured by the area under the curve of the t-distribution corresponding to the t-score. It refers to the probability of obtaining the t-score or a more extreme one when assuming the null hypothesis is true. Thus, to recover the t-score from a two-tailed p-value, we can use the inverse cumulative distribution function to recover the t-score that gives an area under the curve matching the p-value. As such, only the p-value and a degree of freedom are needed.

Once the t-score is recovered, Equation 2 is used to calculate the CI. One caveat of this approach is when p-values are reported using inequalities, although this practise is declining [8]. In these cases, the CI will be larger and more likely to cause type 2 errors. Thus, it is often preferable to use the t-score when reported (section 4.1.1).

4.1.3 From independent means and standard deviations. Depending on whether the study design is within or between subject, the group means (and t-test) will be dependent or independent. Given the means and standard deviations of two between subject groups, an independent t-test can be calculated to obtain a t-score. Equation 2 can then be used to obtain the CI. Note that the same does not apply to a dependent t-test for a within subject condition because the required standard deviation of the paired differences cannot be estimated from means and standard deviations alone. In this case, one of the solutions above should be used.

4.1.4 From a CI at a different confidence level. Given a confidence interval at confidence level $\alpha 0$, we can adapt equation 2 to calculate the CI at $\alpha 1$:

$$MoE_{\alpha 1} = \frac{MoE_{\alpha 0}}{t_{\alpha 0}(df)} \times t_{\alpha 1}(df) \tag{3}$$

4.1.5 From CIs on separate means. Given two CIs on means of two independent groups, the CI of the difference can be recovered by first calculating the mean and standard deviation of each group, and then using the method from section 4.1.3. Assuming the t or normal distribution were used, the CI is symmetrical such that its centre is the mean of the sample. For the standard deviation sd, it can be recovered by rearranging the equation that calculates the MoE (half of the CI),

$$sd = \frac{MoE_\alpha}{t_\alpha(N-1)} \times \sqrt{N} \tag{4}$$

In case another method such as bootstrapping was used to calculate the CI, the recovered sd will be approximate, although it is reasonable to calculate the CI on the difference of means (section 5.5).

Note also that, for two dependent groups, the separate CIs are not enough and the t-score or p-value will be needed to recover the standard deviation of the paired differences.

4.2 Converting to p-values

Given a CI on the mean difference of two groups, and the means of both groups M_1 and M_2, equation 2 can be rearranged to recover the t-score,

$$tscore = \frac{t_\alpha(df)}{MoE_\alpha} \times (M_2 - M_1) \tag{5}$$

Then, the t-score is converted into a two-tailed p-value using the the cumulative distribution function of the t-distribution.

Note that this assumes the t distribution was used to calculate the CI. The validity of this conversion with bootstrapped CIs is evaluated in section 5.5.

4.3 Converting to Standardized Effect Sizes

Whereas section 4.1 used the mean difference as an "unstandardized" effect size, standardized effect sizes such as Cohen's d might be of interest to readers when comparing results that are on different scales and from different experiments [20]. Many effect sizes have been proposed, but most can be converted from one to another. Thus, below we show how to obtain Cohen's d, and how to convert it to a different effect size. These equations can be trivially extended to the case where only CIs are available by first converting the CI to a p-value (section 4.2).

4.3.1 From standard deviations. Cohen's d is calculated by dividing the mean difference by a 'standardizer' which differs depending on the study design. For a between-subject study, the recommended standardizer is the pooled standard deviation [21, 66]. Given the standard deviations of two groups SD_1 and SD_2,

$$d_{\text{between}} = \frac{M_2 - M_1}{\sqrt{\frac{(N_1-1)SD_1^2 + (N_2-1)SD_2^2}{df}}} \tag{6}$$

For a within subject design, the standardizer is usually a pooled average of the standard deviations [21, p 204],

$$d_{\text{within}} = \frac{M_2 - M_1}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}} \tag{7}$$

4.3.2 From a t-score (or p-value). Cohen's d can be obtained from just a t-score (and, by extension, a p-value). We report the equations obtained from Daniël Lakens [66]:

$$d_{\text{between}} = tscore\sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \tag{8}$$

$$d_{\text{within}} = \frac{tscore}{\sqrt{N}} \tag{9}$$

4.3.3 Converting between effect sizes. The conversions between effect sizes have been well covered by previous work, especially in the context of meta-analyses. From Cohen's d, it is possible to calculate the point-biserial correlation and Hedges' g [19]; odds ratios [11]; and the "common language effect size", also referred to as the "probability of superiority" [29].

4.4 Other Considerations

4.4.1 Multiple comparisons and corrected p-values. If p-values are corrected, then a converted CI will also be adjusted which may not reflect the expected confidence level. Corrected p-values often appear in reports using exploratory contrasts to account for multiple comparisons (e.g., post-hoc analysis after an ANOVA). Methods for estimation-based statistics traditionally avoid this issue by planning the analysis a priori to only focus on a few comparisons (typically no more than the degrees of freedom) [21, p 414], or to adjust CIs using different corrections [30]. When the original "uncorrected" CI is desired, it can be obtained either by not relying on the p-value (e.g., using the t-score or standard deviation), or by "unadjusting" the p-value, assuming the original correction method is known.

The goal is to recover the uncorrected p-value p given the corrected p-value p^*. For a Bonferroni correction [75], $p = p^*/n$ where n is the number of pairwise comparisons done. For a Šidák correction [90], $p = 1-(1-p^*)^{1/n}$. For a Holm-Bonferroni correction [52], $p = p^*/(n-i)$ where i is the position of the p-value in the list of sorted p-values of all pairwise comparisons. In some cases, the recovered uncorrected p-values might be inexact: first, a correction for multiplicity typically increases p-values and might make them exceed 1 in which case the p-value is often rounded to 1 and some precision is lost. Second, for Holm-Bonferroni, the rank cannot be recovered when the correction changed the ordering of the p-values. These issues are investigated in detail in section 5.4.

4.4.2 t-test Variations. The equations above apply to common variations of t-tests including those for unequal variances, and unequal sample sizes. For example, for Yuen and Welch's t-test, the degrees of freedom will be different, but the equation to convert the p-value to CI will remain the same, given the correct df is used. Additionally, in some situations, different tests are equivalent to t-tests. For example, the result of an ANOVA on a condition with two-levels will be identical to a t-test, and the t-score is the square root of the F-score [49, 50].

4.4.3 Non-parametric statistics. Tests such as Wilcoxon signed rank and Mann-Whitney are typically reported when data does not follow the assumptions of a t-test. The equations above cannot be used with these tests and it is unclear how CIs could be recovered without the underlying data to find its distribution, and without the test giving an indication of what that distribution might be.

4.4.4 Conversions With Incorrect Reports. One might also wonder what the conversion would do in cases where a t-test was applied on data that clearly breaks t-test assumptions. First, it is important to recognize that a t-test might still be a reasonable choice: for example, the central limit theorem states that, for some data distribution, the mean of the data will be normally distributed given a large enough number of samples, and thus, a t-test could be used. This might explain why t-tests are so prevalent in CHI papers (61% of papers that mentioned a statistical test). But more importantly, the equations are meant to convert the results, not to fix them. Little can be done if a statistical report uses the wrong test and obtains potentially erroneous results.

4.4.5 Chaining Equations. Most equations can be rearranged to calculate the measurements they involve. For the sake of brevity, we presented each equation only once. However, we provide an open source JavaScript library with more than 50 equations and possible rearrangements[5]. Given a set of measurements, the library will iteratively calculate all possible values. For a given value of interest, the library can also describe the possible ways to calculate it, and what values would be needed. It can also identify inconsistencies when there are multiple ways of calculating a value, but they yield different results (using a relative error threshold, currently 0.1).

5 CONVERSION ACCURACY

While the equations in section 4 are exact, written reports often round numbers, may use small samples, and calculate values using methods that could impact the conversion accuracy. Consistent with validation approaches used in the statistics literature [35, 100], we conduct Monte-Carlo simulations of common statistical reports. Our three experiments use the conversion equations above to test the accuracy of: (1) converting reports of t-tests to CI; (2) converting reports of post-hoc pairwise comparisons with corrections to CI; and (3) converting reports of confidence intervals calculated via t-distribution and bootstrap methods to p-values.

The simulations are in python using numpy [47] and pingouin [95] for statistical tests and distributions. The bootstrapped confidence intervals are calculated using arch [89]. Code: https://osf.io/x4ue7.

5.1 Data Generation

As is standard with statistical simulations [35, 100], the datasets used in our fictional reports are automatically generated to test a wide range of study designs. All three experiments use generated designs that are prevalent in HCI with the following shared conditions:

- DESIGN: Either between-subject or within-subject.
- SIZE: The size of each group. We choose to use 8, 12, 24, or 40 because they are the most frequent sample sizes found in HCI studies [12]. However, this does not mean we endorse these small sample sizes since they might result in underpowered studies.
- DECIMALS: The number of decimals used to round all values (mean, standard deviation, t-score, and bounds of CI). For p-values, the rounding is done using the number of significant digits to more closely match what would be reported in a paper. For example, with 1 decimal, a p=0.048 is rounded to 0.05.

Although some experiments might add conditions, they are all performed on *at least* 4 SIZES × 2 DESIGNS × 10,000 repetitions. We use 10,000 repetitions as it has been shown to provide precise approximations with designs typical of HCI studies [86]. The same datasets are re-used when varying DECIMALS. We note cases where experiments add other conditions as appropriate.

[5] http://ns.inria.fr/loki/statslator

5.1.1 Generation Process. The data generation process follows standard practices [35, 100]. For two groups A and B, their samples are randomly drawn from two normal distributions $\mathcal{N}_{\mathcal{A}}(\mu_A, 2)$ and $\mathcal{N}_{\mathcal{B}}(\mu_B, 2)$. When simulating populations with equal means, then $\mu_A = \mu_B = 0$. Otherwise, when simulating populations with different means, $\mu_A \neq \mu_B$ and they are randomly drawn from a standard normal distribution $\mathcal{N}(0, 1)$.

When simulating within-subject designs, a random intercept unique to each subject is added. For example, for a subject X, an intercept α_X is calculated and then added to the values of X in group A and B. To calculate the intercept, we follow the procedure from Elkin et al. [35]. The intercept is randomly drawn from $\mathcal{N}(0, \sigma^2)$ where σ^2 is randomly chosen to be either 0.1, 0.5 or 0.9 to represent a *"reasonable ratio between within-subject variance and between-subject variance"* [35].

To calculate the ground truth statistical power, we calculate t-tests using the data generated with full numerical precision. We then measure the proportion of obtained p-values that are inferior to 0.05 when the simulated populations are truly different. Similarly, to calculate the ground truth coverage, we obtain CIs using the t-distribution of the data will full numerical precision. We then measure the proportion of these CIs that include the true difference in population means.

5.2 Metrics

We report standard metrics used in statistics literature to validate CIs (coverage probability [91]) and p-values (type 1 error rate and power [35]). Recall that our goal is to validate the equations presented in section 4. Thus, we consider a conversion as "correct" if it recovers a p-value or CI that yields a similar score with these metrics as the corresponding p-value or CI calculated from the raw data. Below, we clarify the meaning of each metric.

5.2.1 Coverage of CI. The coverage corresponds to the proportion of calculated CIs that contain the true population mean. For a confidence level of 95%, this proportion should be as close as possible to 95%. This means that over a large number of repetitions, we expect 95% of the 95% CIs to contain the true population mean.

5.2.2 Type I error rate. Given a significance level α, the type 1 error rate is the proportion of false positives where a true null hypothesis is rejected ($p < \alpha$). We use a significance level of 0.05 that is common in HCI, so this proportion should be as close as possible to 5%. Over a large number of repetitions, we expect 5% of p-values to be below 0.05 (even though they are type 1 errors).

5.2.3 Power. Statistical power is the proportion of true positives where a false null hypothesis is correctly rejected ($p < \alpha$). The closer the power is to 100%, the more statistically powerful the test. In practice, power can be much lower, especially when the sample size is small.

5.3 Experiment 1: t-test reports to CIs

This first experiment simulates conversions to obtain 95% CIs from reports comparing two groups with a t-test. In total, 240,000 reports are simulated (4 SIZES × 2 DESIGNS × DECIMALS × 10,000 repetitions). For each report, the 95% CI is calculated using either t-score and means (section 4.1.1); p-value and means (section 4.1.2); or means and standard deviations (section 4.1.3). As baseline, we report the coverage of the 95% CI calculated using the t-distribution of the sample raw data.

Table 1: Mean coverage (and standard deviations) of the 95% CIs calculated from different conversion equations and for different DESIGNS, and DECIMALS. The closest the values are to the one obtained from raw data, the better.

Values Used	DESIGN	DECIMALS		
		1	2	3
t-score + means	within	.918 (.27)	.947 (.22)	.950 (.22)
p-value + means	within	.927 (.26)	.948 (.22)	.950 (.22)
raw data	within	.950 (.22)	.950 (.22)	.950 (.22)
t-score + means	between	.917 (.28)	.947 (.22)	.950 (.22)
p-value + means	between	.927 (.26)	.948 (.22)	.950 (.22)
means + stds	between	.904 (.29)	.938 (.24)	.942 (.23)
raw data	between	.950 (.22)	.950 (.22)	.950 (.22)

5.3.1 Results. On average, three decimals is enough, two decimals give reasonable estimates, but one decimal yields narrower and overoptimistic CIs that do not capture 95% coverage. Table 1 shows coverage of CIs depending on conversion method, the study design, and the numeric precision of the values used. Perhaps because rounding errors propagated, "means + stds" produced the least accurate results with a coverage of 93.8% at two decimals and 90.4% at one decimal.

5.4 Experiment 2: Corrected p-values to CIs

This experiment simulates reports of multiple pairwise t-test comparisons as would be done post-hoc after an omnibus test, such as ANOVA. It differs from experiment 1 in that one report may contain 3, 6, or 10 COMPARISONS (corresponding to an independent variable with either 3, 4, or 5 levels) and the p-values are adjusted to counteract the multiple comparisons problem. The goal is to evaluate the impact of these corrections on the calculated CI, and test the approximations to "unadjust" them. In total, 2,160,000 reports are simulated (3 COMPARISONS × 3 CORRECTIONS × 4 SIZES × 2 DESIGNS × 3 DECIMALS × 10,000 repetitions). The CORRECTION applied to the p-values is either Bonferroni [75], Holm-Bonferroni [52] (that we refer to as Holm to avoid confusion), or Šidák [90]. As baseline comparison, we report coverage for 95% CI calculated using the t-distribution of the sample raw data.

5.4.1 Results. On average, corrected p-values tend to increase the confidence level of the CI to match a 99% CI. For reasons mentioned in section 4.4.1, reversing a correction is approximate and works best for Šidák. For Holm, the reversed correction results in recovering lower CIs (down to 92% confidence) and recovering the CI from the t-score should be preferred. Table 2 reports the breakdown of coverage for the different corrections and uncorrected p-values.

5.5 Experiment 3: CIs to p-values

This experiment simulates reports that include 95% CI and for which we would like to recover p-values. In total, 720,000 reports are simulated (3 CI METHOD × 4 SIZE × 2 DESIGN × 3 DECIMALS × 10,000

Table 2: Mean coverage (and standard deviations) of the 95% CIs calculated from p-values adjusted using a different CORRECTION. For uncorrected values, the closest the values to the the one obtained from raw data, the better.

CORRECTION	DECIMALS		
	1	2	3
Bonferroni	.968 (.18)	.988 (.11)	.990 (.10)
Šidák	.968 (.18)	.988 (.11)	.990 (.10)
Holm	.964 (.19)	.984 (.12)	.986 (.12)
Bonferroni uncorrected	.938 (.24)	.957 (.20)	.959 (.20)
Šidák uncorrected	.931 (.25)	.951 (.22)	.952 (.21)
Holm uncorrected	.916 (.28)	.925 (.26)	.925 (.26)
raw data	.950 (.22)	.950 (.22)	.950 (.22)

Table 3: Mean type 1 error rate (and standard deviations) of the p-values calculated from the 95% CI obtained using different CI METHODS and varying DECIMALS precision. The closest the values to the the one obtained from raw data, the better. Coverage of the 95% CI provided for reference.

CI-METHOD	Coverage	DECIMALS		
		1	2	3
t-CI	.948 (.22)	.047 (.21)	.051 (.22)	.051 (.22)
percentile-CI	.935 (.25)	.078 (.27)	.083 (.28)	.083 (.28)
studentized-CI	.944 (.23)	.049 (.22)	.055 (.23)	.055 (.23)
bCA-CI	.935 (.25)	.075 (.26)	.08 (.27)	.081 (.27)
raw data	-	.051 (.22)	.051 (.22)	.051 (.22)

Table 4: Mean power (and standard deviations) of the p-values calculated from the 95% CI obtained using different CI METHODS and varying DECIMALS precision. The closest the values to the the one obtained from raw data, the better. Coverage of the 95% CI provided for reference.

CI-METHOD	Coverage	DECIMALS		
		1	2	3
t-CI	.948 (.22)	.300 (.46)	.307 (.46)	.307 (.46)
percentile-CI	.935 (.25)	.353 (.48)	.36 (.48)	.361 (.48)
studentized-CI	.944 (.23)	.295 (.46)	.306 (.46)	.307 (.46)
bCA-CI	.935 (.25)	.349 (.48)	.356 (.48)	.357 (.48)
raw data	-	.307 (.46)	.307 (.46)	.307 (.46)

repetitions). The CI METHOD to calculate the 95% CI is either the t-distribution (referred to as t-CI), or popular bootstrapping methods such as percentile CI [32] (percentile-CI), the studentized CI [26] (studentized-CI), or the bias-corrected and accelerated CI [33] (BCa-CI). Bootstrapping methods use 2,000 resamples. As baseline, the results for the p-value obtained from appropriate Student's t-tests are reported.

5.5.1 Results. Overall, p-values recovered from BCa and percentile bootstrapped CIs tend to inflate the type 1 error rate, but are statistically more powerful. This increased number of type 1 errors might be explained by these methods generating CIs with a coverage that does not match 95%. This finding is consistent with previous work that found the percentile and BCa method to perform poorly given small samples (N<50), whereas at this size the t-distrubtion or the studentized bootstrap is usually best [34, 100]. For comparison, our experiments use sample sizes between 8 and 40.

P-values recovered from CI calculated from a t-distribution and studentized bootstrap tend to match the p-values that would have been obtained had a t-test been run, even at low one-decimal numeric precision. Table 3 shows the breakdown of type 1 error rate given the different CI methods and number of decimals. Table 4 shows the same breakdown for statistical power. Note that power may appear low, but this is consistent with what is expected, and has been shown before for such small sample sizes [25].

6 STATSLATOR PDF VIEWER

We developed the Statslator PDF viewer for readers to interactively translate between statistical reporting presentation styles in existing documents and generate statistical complementary presentations like effect sizes and graphical charts. The tool can help readers interpret statistical reports. The user interface was designed to make the capabilities of the conversion equations in our library transparent to the reader, so they are aware of the provenance of the data, what calculations are done, and the quality of conversions.

Readers use Statslator to open and view a PDF document, then select content with statistical reporting they wish to translate into a new presentation style or complementary presentation. The new presentations appear in a sidebar, are highly configurable, and present related statistical values to explore and validate, or values that are easier to interpret correctly such as the common language effect size [62] and S-values [84]. The reader can also choose from different representations, such as animated hypothetical outcome plots [56] and interactive plots, which, despite being powerful representations, are unlikely to be found in existing documents because of publishing formats and workflows. We describe the tool and its features in more detail using three use case scenarios.

6.1 Changing the Style of Statistical Report

Sam got a new smartwatch and decides to review the literature to find the best text entry method for this device. Sam first stumbles across "WatchWriter" [42], an article describing a keyboard for smartwatches. The article reports a user study that compares two ways of operating the keyboard either through taps or gestures. At first glance, it appears that the gesture version is preferable. The article reports *"A one-way between-subjects ANOVA did not show that the overall effect of gesture compared to tap on WPM was significant (F(1, 36) = 1.59, p = 0.21). It did show that the effect on CER and KSPC was significant (F(1, 36) = 4.49, p = 0.04 and F(1, 36) = 248.60, p < 0.001)."*[6]. However, Sam is not very familiar with NHST, and the article does not report effect sizes making it difficult to know if it is worth investing time to learn the gesture technique. Thus, Sam selects the text mentioning the statistical results and the table that reports the means for each condition.

Automatic Extraction & Verification. After the selection, the panel on the right is updated to display two tables filled with statistical

[6]The article mentions a one-way ANOVA with two levels which is strictly equivalent to a t-test and thus is supported by our tool.

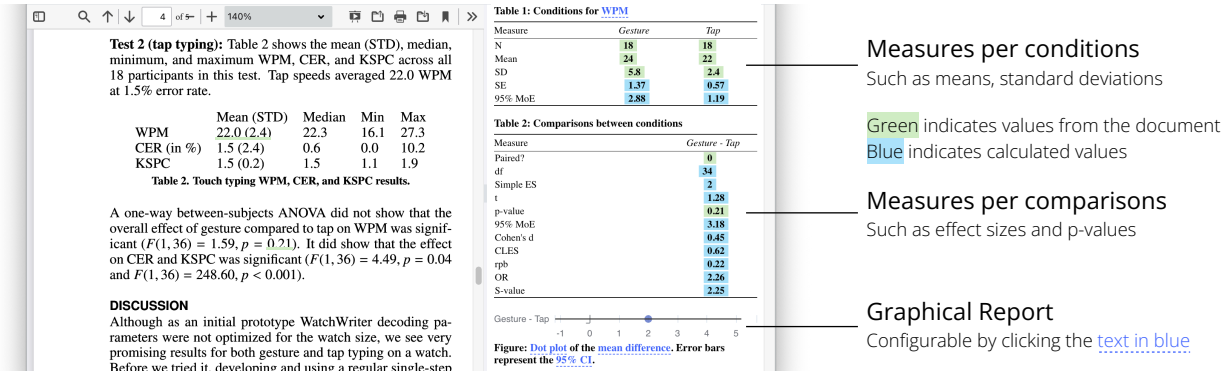
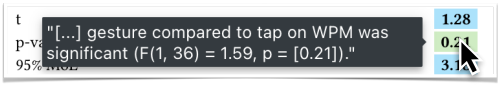
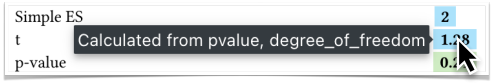


Figure 3: Statslator user interface after selecting a paragraph with statistical information from WatchWriter [42]. The panel on the right shows the statistical measures extracted and calculated, as well as a configurable graphical report.

values from the selection or calculated, as well as a figure of the comparison (figure 3). The background of the cells are coloured based on the provenance of the information: green indicates that the value was obtained from the text or entered manually; and blue indicates that the value was converted. Sam decides to verify that the extracted data is correct by hovering over each value to obtain detailed information. When hovering over a value obtained from the text, a tooltip shows the sentence where the value was extracted from and the corresponding value in the PDF is underlined.

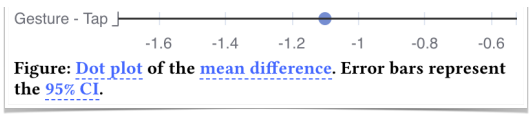


When hovering over a value that was converted by our equations, a tooltip shows details of the calculation. The most accurate conversions are prioritized. For example, for the CI the t-score and means are used even though the standard deviations, p-values are also available.

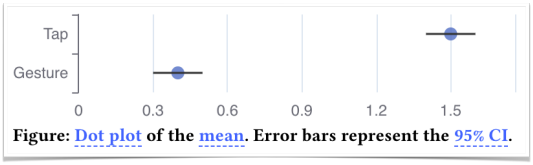


Switching between Dependent Variables. By default, the generated report shows the results for words per minute (WPM). Sam clicks the blue underlined text in the caption of the first table in the panel to change the dependent variable and show the KSPC. Sam realizes that Cohen's d for the comparison of KSPC is quite large (=5.5) compared to the other dependent variables compared. Sam cannot recall the interpretation of Cohen's d, but the tool shows the common language effect size (=.99) that Sam knows to interpret as "when gesture was compared to tap, in 99 of 100 pairs gesture had a lower KSPC than tap."

Configuration of the Graphical Statistical Report. Similarly, the figure generated by the system is interactive and shows the 95% confidence interval on the mean difference.



Sam knows about the overlapping rule for the 95% CIs of independent means and decides to change the figure by clicking the blue underlined text in the caption from "mean difference" to "means".

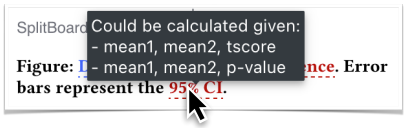


However, after some tinkering, Sam is not sure anymore of the correct interpretation of a 95% confidence interval. Instead, Sam switches the chart to an animated hypothetical outcome plot [56] which, after a few seconds, helps Sam gain intuition for the distribution of the data.

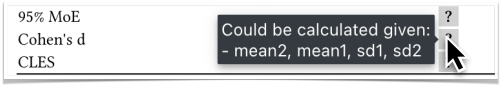
6.2 Comparing Two Reports

After further inspection, Sam realizes that WatchWriter relies on a statistical decoder which makes it difficult to enter words that are out-of-vocabulary. Instead, Sam investigates two alternative techniques that support OOV: SplitBoard [53] and Swipeboard [16]. While both articles present user studies, the two techniques are not compared with each other. Worse, the two studies use different protocols and different study durations, and Sam decides that judgment solely based on the reported means might be misleading. However, both studies include a comparison to ZoomBoard [78], a third text entry technique. Thus, while keeping in mind that the two studies differ in many ways, Sam decides to calculate the standardized effect sizes for the two techniques compared to ZoomBoard. Sam had already opened the papers in Statslator, so Sam begins by selecting the paragraphs containing the statistical results in both papers.

Helping the System with Missing Values. When extracting data from SplitBoard, Sam notices that links for some values are red indicating that they are not available.



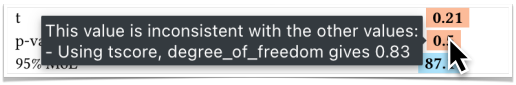
Sam hovers over the missing values in the table to get an indication of what information would be needed for that value to be calculated.



When hovering over Cohen's d, the system indicates that the means and standard deviation are needed to perform the conversion. However, the SplitBoard paper does not report aggregated means with standard deviations. Sam decides to retrieve these values from the line chart of the WPM using an accurate chart data extraction tool [58, 73]. For SplitBoard, the t-score was directly obtained enabling the calculation of Cohen's d. Once the missing values are added in the table, the other values are calculated and Sam can review the effect sizes: Cohen's $d = 0.64$ for SwipeBoard, and $d = 7.27$ for SplitBoard.

6.3 Checking Correctness

While browsing recent preprints, Sam finds a brand new text entry method that looks promising. As always, Sam starts by selecting the statistical report to display the results in a different style. This time, the p-value and t-score have a red background: the system detected an inconsistency after cross-checking different ways values can be obtained.



It appears that the t-score does not match with the reported p-values. Sam knows how these mistakes can easily occur when writing papers [69, 77] and decides to send an email to the authors to warn them.

6.4 Implementation

Statslator is implemented using TypeScript using React [85] with PrimeReact [82] for the interface; PDF.js [80] for the PDF viewer; and ECharts [31] to generate dynamic visualizations. The source code is available[7]. Below, we detail the implementation of the text extraction to recover statistical values.

6.4.1 Extracting Statistical Information. Upon selecting a text in a document, the text is extracted and sent to OpenAI's GPT-3.5 through the official Chat API[8]. GPT-3.5 takes care of extracting the statistical information contained in the selection, even if the text contains tables and complex sentence structures. GPT-3.5 is a large language model based on a transformer architecture [96] and powers ChatGPT. The GPT models are state-of-the-art in many natural language processing tasks, especially for complex sentence structures [83]. The task of extracting statistical information is no exception: in our tests, GPT-3.5 outperformed all alternatives. GPT-3.5 requires a hand-crafted *prompt* that explains the task. We engineered the following prompt through repeated experiments:

```
<excerpt from the paper>
Answer with this JSON structure:
{"conditions": [/* reported numbers that refer to a condition following this JSON
      format: [<number>, <type> /* example: mean, sd, upper CI */, <condition
      >]*/],
"comparisons": [/* reported numbers that refer to a t-test comparison following this
      JSON format: [<number>, <type> /* example: p-value, t-score */, <
      condition1>, <condition2>*/]]}
```

We use *"gpt-3.5-turbo"* which has an input limit of about 3,000 words, forbidding long text selections[9]

For privacy reasons or because it is a paid service, we also support alternatives to GPT-3.5. First, readers can always input the numbers manually. Second, we also provide an extraction algorithm that relies on regular expressions. The algorithm searches for APA symbols such as "M=" to extract statistical values (similar to section 3). We group values based on their order of appearance. For example, the first mentioned mean and standard deviation are grouped as a single condition. Similarly, the first p-value is associated with the first comparison of the two conditions mentioned.

7 DISCUSSION

Our work highlights three aspects: (1) even though most papers report statistics in a specific way, they usually contain enough information to convert them into a different statistical reporting style; (2) most reporting practices are compatible with accurate conversions; and (3) a PDF viewer that embeds these conversions can enable readers to control the presentation of statistics in existing documents to better understand the results, compare documents, and verify correctness. Before outlining the limitations and future work, it is important to clarify what this work is not.

In no circumstances does this work replace the proper practice of statistics by authors, nor does it weaken arguments for statistical reforms such as using CIs instead of p-values [20]. The differences between an NHST and estimation approach are more than just the presentation of the results. For example, estimation-based thinking also implies a different way of formulating research questions and drawing conclusions which can hardly be an afterthought [14]. And best practices involve planning studies and following open science procedures [18, 21, 27]. Instead, our work helps *readers* desiring a different presentation of statistical results, perhaps to draw their own conclusions. If anything, our hope is that showing that conversions are possible will motivate authors to choose the method most appropriate to them and their research questions [67] without worrying about possible push-back. As expressed by Andy Cockburn in response to an alt.chi article encouraging authors to use estimation and avoid dichotomous reports: *"Sometimes, however, the author would prefer to NOT report dichotomous outcomes (for good reasons), but is compelled to do so by their fear/knowledge that if not included, reviewers will expect it and criticise its absence"* [8].

7.1 Limitations

7.1.1 The analyses might have missed some papers and study designs. With large-scale experiments, it is difficult to consider all cases. In section 3, when quantifying the proportions of each report at CHI, some papers might have fallen through the cracks, either because

[7] http://ns.inria.fr/loki/statslator
[8] https://platform.openai.com/docs/api-reference
[9] The limit has been increased with newer versions of the GPT models. Full papers could now be parsed directly.

they use complex sentence structures, or because the reports are done in figures and tables. Our analysis was meant to motivate the feasibility of conversions and have reference proportions to decide the most appropriate input values for the equations.

Similarly, there are an infinite number of parameters and study designs that could be tested in the experiment section 5, but we chose to focus on those most prevalent in HCI. This had the side-effect of steering our experiments and equations towards the use of small samples which have been covered relatively poorly in the past. However, as the sample size increases, some of these considerations are no more relevant considering the central limit theorem [36].

7.1.2 The conversions might fail in some situations. As a corollary of the infinite space of study designs, we cannot guarantee that the conversion will be correct in all cases. Most statistical tests have underlying assumptions that might in fact be violated by the data and yield unexpected results. Generally, our conversion equations are based on the assumption that the authors respected the assumptions of the methods they used. However, it is not uncommon for scientists to use an inappropriate test [69, 99] or make mistakes [76, 77], especially considering the challenges associated with statistics in HCI [61]. In these cases, the results will most likely be as incorrect as their original presentation.

7.1.3 The text extraction of statistical values might fail. We extract statistical values using a large language model (GPT-3.5). While we found it to work well in our limited tests, we did not formally assess its accuracy. Instead, we focused on making sure readers could verify the extraction was successful through provenance verification features such as highlighting the value in the document and showing the sentence that contained the value on hover.

7.2 Future Work

7.2.1 Support for more papers through access to more data. The main hurdle in converting a statistical report is the lack of data. We focused on t-tests and papers reporting means and p-values or t-scores because we found these to be the most common at CHI. Conversion between any statistical report is theoretically possible given access to the raw data. Thus, one extension of our work could be to leverage that data when available. For example, the system could automatically pull data from a repository such as https://osf.io/. However, "Open Data" has still a long way to go in communities such as CHI where less than 1% of papers make data available [1, 73]. Other times, the data is in the document but buried inside data visualizations. In these cases, tools to "reverse-engineer" data visualizations (e.g. [58, 73]) could make the system work with a broader set of scientific articles.

7.2.2 Support for other conversions. Future work could support more tests and conversions towards a broader set of reporting styles. Tests such as Wilcoxon signed-rank and Mann-Whitney U would be a natural extension given their prevalence. The challenge is to infer the underlying distribution of the data. When the data is available, this could be done through visual inspection or by finding the best-fitting distribution. Otherwise, the distribution could be inferred from the type of measure. Additionally, effect sizes could be further supported for different tests, including η^2 for ANOVAs. And a CI on these effect sizes could be calculated given limited data using approaches such as the noncentrality parameter [23, 93].

In terms of statistical reports, we focused on NHST and estimation because they are fuelling many debates within the scientific community [15, 27]. However, other approaches such as Bayesian statistics could be supported. For example, a Bayesian t-test has been proposed and can be calculated given a t-score and a prior [87]. The prior could be controlled by readers to reflect their optimism and knowledge, similar to what Dragicevic et al. proposed [28].

7.2.3 Statistical linting, meta-analyses, and statistical education. There are many use cases that could be derived from our system. First, our focus was on readers, but the mechanisms leveraged to detect inconsistencies could power a statistical linter for authors. Similar to *statscheck* [76] that detects inconsistencies between the reported p-value and reported test statistic, a statistical linter could leverage our system to cross-check the different ways of obtaining a value. This would allow the detection of serious problems such as using a t-test that does not match the study design.

Similarly, many of our equations could be useful in meta-analyses, especially within fields like HCI that deal with small samples for which typical meta-analyses practices are overoptimistic [3, 4]. Our system could help scientists recover accurate data into a customizable and standardized statistical measure.

Finally, we assumed readers have some experience and preferences regarding statistics, but our PDF viewer could also be used as an educational tool. Transitioning between statistical representations is particularly useful to develop an intuition [21, 24]. Akin to explorable multiverse analyses reports, some options could be educational to give readers a better grasp of certain concepts [28].

8 CONCLUSION

While much of the debate around statistics has been focused on how authors should practise and report them, little has been done to support readers and the thousands of documents already published. Through theoretical and empirical evidence, we showed that a majority of CHI papers report enough information to be converted to different statistical reporting styles and that the conversions remain mostly accurate under common reporting practices. We also describe the design and implementation of a PDF viewer to turn existing papers into the statistical reporting style readers prefer. Our hope is to provide an immediate solution to reconciliate readers with statistical reports, all while unburdening authors to let them focus on proper statistical practices.

ACKNOWLEDGMENTS

This work was made possible by NSERC Discovery Grant 2018-05187 and the LAI Réapp.

REFERENCES

[1] Jacob Abbott, Haley MacLeod, Novia Nurain, Gustave Ekobe, and Sameer Patil. 2019. Local Standards for Anonymization Practices in Health, Wellness, Accessibility, and Aging Research at CHI. 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/10.1145/3290605.3300692

[2] Douglas G. Altman and J. Martin Bland. 1996. Statistics Notes: Presentation of Numerical Data. *BMJ* 312, 7030 (March 1996), 572. https://doi.org/10.1136/bmj.312.7030.572

[3] Douglas G. Altman and J. Martin Bland. 2011. How to Obtain the Confidence Interval from a P Value. *BMJ* 343 (Aug. 2011), d2090. https://doi.org/10.1136/bmj.d2090

[4] Douglas G. Altman and J. Martin Bland. 2011. How to Obtain the P Value from a Confidence Interval. *BMJ* 343 (Aug. 2011), d2304. https://doi.org/10.1136/bmj.d2304

[5] American Psychologic Association. 2020. *Publication Manual of the American Psychological Association* (6th edition ed.). American Psychological Association, Washington, DC.

[6] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 661–671. https://doi.org/10.1109/TVCG.2018.2865119

[7] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods* 10, 4 (2005), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

[8] 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

[9] Lonni Besançon, Yvonne Jansen, Andy Cockburn, and Pierre Dragicevic. 2021. *Definitely Maybe: Hedges And Boosters in the HCI Literature*. Preprint. Open Science Framework. https://doi.org/10.31219/osf.io/mjg7h

[10] Dennis D. Boos and Leonard A. Stefanski. 2011. P-Value Precision and Reproducibility. *The American Statistician* 65, 4 (Nov. 2011), 213–221. https://doi.org/10.1198/tas.2011.10129

[11] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis* (1st edition ed.). Wiley, Chichester, U.K.

[12] 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

[13] Alberto Cairo. 2019. *How Charts Lie: Getting Smarter about Visual Information* (illustrated edition ed.). WW Norton, New York.

[14] Robert J. Calin-Jageman and Geoff Cumming. 2019. Estimation for Better Inference in Neuroscience. *eNeuro* 6, 4 (2019), ENEURO.0205–19.2019. https://doi.org/10.1523/ENEURO.0205-19.2019

[15] Robert J. Calin-Jageman and Geoff Cumming. 2019. The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. *The American Statistician* 73, sup1 (March 2019), 271–280. https://doi.org/10.1080/00031305.2018.1518266

[16] Xiang 'Anthony' Chen, Tovi Grossman, and George Fitzmaurice. 2014. Swipeboard: A Text Entry Technique for Ultra-Small Interfaces That Supports Novice to Expert Transitions. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, Honolulu Hawaii USA, 615–620. https://doi.org/10.1145/2642918.2647354

[17] William S. Cleveland. 2004. *The Elements of Graphing Data* (2 edition ed.). Hobart Press, Murray Hill, N.J.

[18] 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 (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173715

[19] Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine (Eds.). 2019. *The Handbook of Research Synthesis and Meta-Analysis* (3rd edition ed.). Russell Sage Foundation, New York.

[20] Geoff Cumming. 2014. The New Statistics: Why and How. *Psychological Science* 25, 1 (Jan. 2014), 7–29. https://doi.org/10.1177/0956797613504966

[21] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the New Statistics: Estimation, Open Science, and Beyond* (1st edition ed.). Routledge, London : New York.

[22] Geoff Cumming, Fiona Fidler, and David L. Vaux. 2007. Error Bars in Experimental Biology. *Journal of Cell Biology* 177, 1 (April 2007), 7–11. https://doi.org/10.1083/jcb.200611141

[23] Geoff Cumming and Sue Finch. 2001. A Primer on the Understanding, Use, and Calculation of Confidence Intervals That Are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement* 61, 4 (Aug. 2001), 532–574. https://doi.org/10.1177/0013164401614002

[24] Geoff Cumming, Mark Zangari, and Neil Thomason. 1995. Designing Software for Cognitive Change: StatPlay and Understanding Statistics. In *World Conference on Computers in Education VI: WCCE '95 Liberating the Learner, Proceedings of the Sixth IFIP World Conference on Computers in Education, 1995*, J. David Tinsley and Tom J. van Weert (Eds.). Springer US, Boston, MA, 753–765. https://doi.org/10.1007/978-0-387-34844-5_71

[25] Marie Delacre, Daniël Lakens, and Christophe Leys. 2017. Why Psychologists Should by Default Use Welch's t-Test Instead of Student's t-Test. *International Review of Social Psychology* 30, 1 (April 2017), 92. https://doi.org/10.5334/irsp.82

[26] Thomas J. DiCiccio and Bradley Efron. 1996. Bootstrap Confidence Intervals. *Statist. Sci.* 11, 3 (Sept. 1996), 189–228. https://doi.org/10.1214/ss/1032280214

[27] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[28] 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

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

[30] Olive Jean Dunn. 1961. Multiple Comparisons among Means. *J. Amer. Statist. Assoc.* 56, 293 (March 1961), 52–64. https://doi.org/10.1080/01621459.1961.10482090

[31] Apache 2015. *Apache ECharts*. Apache. https://echarts.apache.org/en/index.html

[32] Bradley Efron. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial & Applied Mathematics,U.S., Philadelphia, Pa.

[33] Bradley Efron. 1987. Better Bootstrap Confidence Intervals. *J. Amer. Statist. Assoc.* 82, 397 (March 1987), 171–185. https://doi.org/10.1080/01621459.1987.10478410

[34] Bradley Efron. 2000. The Bootstrap and Modern Statistics. *J. Amer. Statist. Assoc.* 95, 452 (2000), 1293–1296. https://doi.org/10.2307/2669773 jstor:2669773

[35] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[36] Hans Fischer. 2011. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-87857-7

[37] Tong Gao, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3005–3014. https://doi.org/10.1145/2556288.2557228

[38] Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. 2002. Let's Practice What We Preach. *The American Statistician* 56, 2 (May 2002), 121–130. https://doi.org/10.1198/000313002317572790

[39] David Giofrè, Ingrid Boedker, Geoff Cumming, Carlotta Rivella, and Patrizio Tressoldi. 2022. The Influence of Journal Submission Guidelines on Authors' Reporting of Statistics and Use of Open Research Practices: Five Years Later. *Behavior Research Methods* (Oct. 2022). https://doi.org/10.3758/s13428-022-01993-3

[40] David Giofrè, Geoff Cumming, Luca Fresc, Ingrid Boedker, and Patrizio Tressoldi. 2017. The Influence of Journal Submission Guidelines on Authors' Reporting of Statistics and Use of Open Research Practices. *PLOS ONE* 12, 4 (April 2017), e0175583. https://doi.org/10.1371/journal.pone.0175583

[41] Steven Goodman. 2008. A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology* 45, 3 (July 2008), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

[42] Mitchell Gordon, Tom Ouyang, and Shumin Zhai. 2016. WatchWriter: Tap and Gesture Typing on a Smartwatch Miniature Keyboard with Statistical Decoding. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 3817–3821. https://doi.org/10.1145/2858036.2858242

[43] Transparent Statistics in Human-Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. https://doi.org/10.5281/zenodo.1186169

[44] Ken Gu, Eunice Jun, and Tim Althoff. 2023. Understanding and Supporting Debugging Workflows in Multiverse Analysis. arXiv:2210.03804 [cs]

[45] Brian D. Hall, Yang Liu, Yvonne Jansen, Pierre Dragicevic, Fanny Chevalier, and Matthew Kay. 2022. A Survey of Tasks and Visualizations in Multiverse Analysis Reports. *Computer Graphics Forum* 41, 1 (2022), 402–426. https://doi.org/10.1111/cgf.14443

[46] Lisa Lavoie Harlow, Stanley A. Mulaik, and James H. Steiger (Eds.). 1997. *What If There Were No Significance Tests?* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. xviii, 446 pages.

[47] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array Programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[48] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besançon. 2021. Can Visualization Alleviate Dichotomous Thinking? Effects of Visual Representations on the Cliff Effect. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (Aug. 2021), 3397–3409. https://doi.org/10.1109/TVCG.2021.3073466 arXiv:2002.07671 [cs, stat]

[49] R.N. Henson. 2015. Analysis of Variance (ANOVA). In *Brain Mapping*. Elsevier, Los Angeles, CA, USA, 477–481. https://doi.org/10.1016/B978-0-12-397025-1.00319-5

[50] Michael H. Herzog, Gregory Francis, and Aaron Clarke. 2019. ANOVA. In *Understanding Statistics and Experimental Design : How to Not Lie with Statistics*, Michael H. Herzog, Gregory Francis, and Aaron Clarke (Eds.). Springer International Publishing, Cham, 67–82. https://doi.org/10.1007/978-3-030-03499-3_6

[51] 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

[52] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. jstor:4615733

[53] Jonggi Hong, Seongkook Heo, Poika Isokoski, and Geehyuk Lee. 2015. SplitBoard: A Simple Split Soft Keyboard for Wristwatch-sized Touch Screens. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1233–1236. https://doi.org/10.1145/2702123.2702273

[54] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013. Contextifier: Automatic Generation of Annotated Stock Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2707–2716. https://doi.org/10.1145/2470654.2481374

[55] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving Comprehension of Measurements Using Concrete Re-expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173608

[56] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE* 10, 11 (Nov. 2015), e0142444. https://doi.org/10.1371/journal.pone.0142444

[57] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A High-level Language and Runtime System for Automating Statistical Analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 591–603. https://doi.org/10.1145/3332165.3347940

[58] Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. 2017. ChartSense: Interactive Data Extraction from Chart Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6706–6717. https://doi.org/10.1145/3025453.3025957

[59] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2019. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 892–902. https://doi.org/10.1109/TVCG.2018.2864909

[60] Jonathan P. Kastellec and Eduardo L. Leoni. 2007. Using Graphs Instead of Tables in Political Science. *Perspectives on Politics* 5, 4 (Dec. 2007), 755–771. https://doi.org/10.1017/S1537592707072209

[61] 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

[62] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting Scientific Results in Perspective: Improving the Communication of Standardized Effect Sizes. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–14. https://doi.org/10.1145/3491102.3502053

[63] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 38–48. https://doi.org/10.1145/2858036.2858440

[64] Rex B. Kline. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. American Psychological Association, Washington, DC, US. xii, 325 pages. https://doi.org/10.1037/10693-000

[65] Martin Krzywinski and Naomi Altman. 2013. Error Bars. *Nature Methods* 10, 10 (Oct. 2013), 921–922. https://doi.org/10.1038/nmeth.2659

[66] Daniël Lakens. 2013. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology* 4 (Nov. 2013), 863. https://doi.org/10.3389/fpsyg.2013.00863

[67] Daniel Lakens. 2019. The Practical Alternative to the P-Value Is the Correctly Used p-Value. https://doi.org/10.31234/osf.io/shm8v

[68] Marie-Paule Lecoutre, Jacques Poitevineau, and Bruno Lecoutre. 2003. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology* 38, 1 (2003), 37–45. https://doi.org/10.1080/00207590244000250

[69] Guangping Liang, Wenliang Fu, and Kaifa Wang. 2019. Analysis of T-Test Misuses and SPSS Operations in Medical Research Papers. *Burns & Trauma* 7 (Jan. 2019), s41038–019–0170–3. https://doi.org/10.1186/s41038-019-0170-3

[70] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

[71] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2023. Charagraph: Interactive Generation of Charts for Realtime Annotation of Data-Rich Paragraphs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581091

[72] Damien Masson, Sylvain Malacria, Edward Lank, and Géry Casiez. 2020. Chameleon: Bringing Interactivity to Static Digital Documents. 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/10.1145/3313831.3376559

[73] Damien Masson, Sylvain Malacria, Daniel Vogel, Edward Lank, and Géry Casiez. 2023. ChartDetective: Easy and Accurate Interactive Data Extraction from Complex Vector Charts. 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/10.1145/3544548.3581113

[74] Blakeley B. McShane and David Gal. 2016. Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence. *Management Science* 62, 6 (June 2016), 1707–1718. https://doi.org/10.1287/mnsc.2015.2212

[75] Rupert G. Miller. 1981. *Simultaneous Statistical Inference*. Springer, New York, NY. https://doi.org/10.1007/978-1-4613-8122-8

[76] Michèle B. Nuijten and Joshua R. Polanin. 2020. "Statcheck": Automatically Detect Statistical Reporting Inconsistencies to Increase Reproducibility of Meta-Analyses. *Research Synthesis Methods* 11, 5 (2020), 574–579. https://doi.org/10.1002/jrsm.1408

[77] Michèle B. Nuijten, Marcel A. L. M. van Assen, C. H. J. Hartgerink, Sacha Epskamp, and Jelte Wicherts. 2017. The Validity of the Tool "Statcheck" in Discovering Statistical Reporting Inconsistencies. https://doi.org/10.31234/osf.io/tcxaj

[78] Stephen Oney, Chris Harrison, Amy Ogan, and Jason Wiese. 2013. ZoomBoard: A Diminutive Qwerty Soft Keyboard Using Iterative Zooming for Ultra-Small Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2799–2802. https://doi.org/10.1145/2470654.2481387

[79] Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How Deceptive Are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1469–1478. https://doi.org/10.1145/2702123.2702608

[80] Mozilla 2011. *PDF.Js*. Mozilla. https://mozilla.github.io/pdf.js/

[81] 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

[82] PrimeTek Informatics 2017. *PrimeReact | React UI Component Library*. PrimeTek Informatics. https://www.primefaces.org/primereact

[83] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

[84] Zad Rafi and Sander Greenland. 2020. Semantic and Cognitive Tools to Aid Statistical Science: Replace Confidence and Significance by Compatibility and Surprise. *BMC Medical Research Methodology* 20, 1 (Sept. 2020), 244. https://doi.org/10.1186/s12874-020-01105-9

[85] Meta 2013. *React – A JavaScript Library for Building User Interfaces*. Meta. https://reactjs.org/

[86] Grégoire Richard, Thomas Pietrzak, Ferran Argelaguet, Anatole Lécuyer, and Géry Casiez. 2022. Within or Between? Comparing Experimental Designs for Virtual Embodiment Studies. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Christchurch, New Zealand, 186–195. https://doi.org/10.1109/VR51125.2022.00037

[87] Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. Bayesian t Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review* 16, 2 (April 2009), 225–237. https://doi.org/10.3758/PBR.16.2.225

[88] Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. 2021. *Multiverse: Multiplexing Alternative Data Analyses in R Notebooks.* Preprint. Open Science Framework. https://doi.org/10.31219/osf.io/yfbwm

[89] Kevin Sheppard, Stanislav Khrapov, Gábor Lipták, mikedeltalima, Rob Capellini, alejandro cermeno, Hugle, esvhd, Snyk bot, Alex Fortin, JPN, Matt Judell, Weiliang Li, Austin Adams, jbrockmendel, M. Rabba, Michael E. Rose, Nikolay Tretyak, Tom Rochette, UNO Leo, Xavier RENE-CORAIL, Xin Du, and Burak Çelik. 2022. *bashtage/arch: Release 5.3.1.* https://doi.org/10.5281/zenodo.6684078

[90] Zbyněk Šidák. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Amer. Statist. Assoc.* 62, 318 (June 1967), 626–633. https://doi.org/10.1080/01621459.1967.10482935

[91] Kurex Sidik and Jeffrey N. Jonkman. 2002. A Simple Confidence Interval for Meta-Analysis. *Statistics in Medicine* 21, 21 (2002), 3153–3159. https://doi.org/10.1002/sim.1262

[92] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (Sept. 2016), 702–712. https://doi.org/10.1177/1745691616658637

[93] James H. Steiger. 20040510. Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis. *Psychological Methods* 9, 2 (20040510), 164. https://doi.org/10.1037/1082-989X.9.2.164

[94] Gail M. Sullivan and Richard Feinn. 2012. Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education* 4, 3 (Sept. 2012), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

[95] Raphael Vallat. 2018. Pingouin: Statistics in Python. *Journal of Open Source Software* 3, 31 (Nov. 2018), 1026. https://doi.org/10.21105/joss.01026

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

[97] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play.* ACM, Virtual Event Canada, 4–18. https://doi.org/10.1145/3410404.3414229

[98] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding Novices in Statistical Analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* Association for Computing Machinery, New York, NY, USA, 2693–2702. https://doi.org/10.1145/2702123.2702347

[99] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11).* Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[100] Yaqian Zhu and John Kolassa. 2018. Assessing and Comparing the Accuracy of Various Bootstrap Methods. *Communications in Statistics - Simulation and Computation* 47, 8 (Sept. 2018), 2436–2453. https://doi.org/10.1080/03610918.2017.1348516