A Meta-Analysis of Effect Sizes of CHI Typing Experiments

Obukhova, Natalia

Abstract: While designing an HCI experiment, planning the sample size with a priori power analysis is often skipped due to the lack of reference effect sizes. On the one hand, it can lead to a false-negative result, missing the effect that is present in the population. On the other hand, it poses a risk of spending more resources if the number of participants is too high. In this work, I present the reference for small, medium, and large effect sizes for typing experiments based on a meta-analysis of well-cited papers from CHI conference. This effect size ruler can be used to conduct a priori power analysis or assess the magnitude of the found effect. This work also includes comparisons to other fields and conclude with a discussion of the existing issues with reporting practices and data availability. This paper and all data and materials are freely available at https://osf.io/nqzpr.

DOI: https://doi.org/10.1145/3411763.3451520

A Meta-Analysis of Effect Sizes of CHI Typing Experiments

Natalia Obukhova
University of Zurich
Zurich, Switzerland
natalia.obukhova@uzh.ch

ABSTRACT

While designing an HCI experiment, planning the sample size with *a priori* power analysis is often skipped due to the lack of reference effect sizes. On the one hand, it can lead to a false-negative result, missing the effect that is present in the population. On the other hand, it poses a risk of spending more resources if the number of participants is too high. In this work, I present the reference for small, medium, and large effect sizes for typing experiments based on a meta-analysis of well-cited papers from CHI conference. This *effect size ruler* can be used to conduct *a priori* power analysis or assess the magnitude of the found effect. This work also includes comparisons to other fields and conclude with a discussion of the existing issues with reporting practices and data availability. This paper and all data and materials are freely available at https://osf.io/nqzpr.

CCS CONCEPTS

• **Human-centered computing** → **Laboratory experiments**; **Text input**.

KEYWORDS

experimental design, controlled experiments, power analysis, text entry, systematic review

ACM Reference Format:
Natalia Obukhova. 2021. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451520

1 BACKGROUND AND MOTIVATION

In the field of Human-Computer Interaction (HCI), one way to validate newly developed interaction techniques is a controlled experiment. Hornbæk et al. [19] did a systematic review of 891 papers between 2008 and 2010 in 5 major HCI venues and found that more than 48% of papers studying interactions with user interfaces report experiments. When designing controlled experiments with human participants, researchers need to decide on a diverse set of parameters, including how many participants should be recruited.

One of the interaction techniques widely researched with experiments is text-entry. Typing is used for information transfer and staying connected with the people around us. Researchers started studying typing methods since the invention of the typewriter in 1868. Hereafter, a typing experiment is referred to as an experiment that involves studying the effect of different keyboards and environments on typing performance metrics. However, to my knowledge, there is no reference for effect sizes for planning typing experiments. MacKenzie and Tanaka-Ishii [29, p. 85] suggest using a number of participants for typing experiments close to "the other research with a similar methodology". However, with the rapid development of technology and interfaces, it is increasingly difficult for researchers to identify relevant previous work that can be used to plan their experiment. For example, between the years 2000 and 2010, researchers have done considerable work on virtual keyboards for touch-sensitive screens of varying sizes. In contrast, most papers in the same field written in 2020 include text-entry in virtual and augmented reality with voice and brain input. Thus, finding the previous work about typing with the same methodology is not a trivial task.

One statistical method for sample size planning is *a priori* power analysis, based on three parameters: α (Type I error probability), $1 - \beta$ (statistical power), and a standardized effect size. A standardized effect size describes the difference between means and standard deviations of two groups. Wang et al. [42, p. 2] discuss the difficulties of estimating such standardized effect size. Cohen [5, p. 24-27] proposed a convention of small, medium, and large effect sizes. However, this convention is based on the difference in human heights and in intelligence quotients. Hence, Cohen advised researchers to create and use local standards for each field rather than blindly following his convention [5, p. 24-27]. In this paper, I refer to effect size conventions of the small, medium, and large effect sizes as *effect size ruler*.[1]

Researchers showed that the local effect size standards in Psychology [27] and Software Engineering [21] differ from Cohen's convention. Kampenes et al. [21] conducted a systematic review of the software engineering experiments. They collected 284 Hedges's g effect sizes and binned them into three categories: small, medium, and large. The authors used the median value in each of the bins to create the effect size ruler. Kampenes et al. compared their results to meta-analyses in Psychology, Education, and Behavioral Sciences [26], as well as the Cohen's convention [5, p. 24-27], and show the clear difference in local standards for these fields ([21, Table 11]).

In the field of HCI, there are several quantitative reviews and meta-analyses. Hornbæk and Law [18] conducted a meta-analysis of 73 studies. The results show that the Pearson's correlations among the usability measures are low, and the authors argued that

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451520

[1] This name is inspired by Cumming's metaphor of standardized effect sizes as the *rubber ruler* [6]. The ruler metaphor indicates two usages of the effect size convention: as a guide for approximating the effect sizes in study planning and as a reference to compare study results.

using only one usability measure might miss important information. Stowell et al. [36] performed a thematic analysis of 84 papers and three meta-analyses across five papers. They studied the impact of mHealth technologies on vulnerable populations in the U.S. In their meta-analyses, they did not find evidence of the success of interventions. Yee et al. [45] performed formal and informal meta-analyses of the impact of human-like faces on user experience of performance and subjective metrics. They found an effect of inclusion on subjective metrics larger than on performance ones and more importance of the inclusion effect than the effect of face's realism. Ma et al. [28] conducted a meta-analysis of 19 papers about virtual humans' effect on medical treatment. The authors found that virtual humans significantly improves health-related outcomes. In summary, these works only provide a reference effect size for the specific topic but none for typing experiments. As there exists no local effect size standard for typing experiments, researchers might default to using Cohen's conventions. However, using these conventions risks not finding significant results even if the effect exists in the population but is too small for the chosen sample size.

To address this problem, I contribute a reference effect sizes ruler for typing experiments based on a meta-analytical approach that enables researchers to plan sample sizes with *a priori* power analysis and compare their results within the field. I provide an additional effect sizes ruler for the common dependent variable "words per minute".

2 METHOD

The goal of this paper is to create an *effect size ruler* for typing experiments in HCI. In this work, I use a multi-level meta-analytical approach to calculate the interval estimate for the small, medium, and large effect sizes [16].[2] A meta-analysis is a statistical procedure that allows researchers to combine the findings from different studies. The benefit of using a meta-analysis instead of taking simple means or medians, which give only the point estimate, is that it allows to calculate interval estimate, and also takes the weight of each study into account (the more participants are in the study, the higher is the weight). A multi-level meta-analysis has the advantage of taking several effect sizes that are present within one study into account [3].

2.1 Studies selection

The goal of the selection process is to identify typing experiments that are likely to be influential in the field. Figure 1 shows the PRISMA diagram that summarizes the selection process [31].

I performed a keyword search on the ACM digital library in all CHI proceedings on July 21 2020 to ensure the first criterion—papers about typing experiments published in CHI. CHI is an umbrella conference for HCI-related topics and therefore yield a diverse set of papers. I used keywords "text input", "text entry", "typing", and "keyboard" to cover for all papers about typing experiments, and the word "experiment" in the full-text. This search yielded 108 articles that were published between 2008 and 2020.

Ideally, this meta-analysis should include all resulting papers. However, due to a lack of standard in statistical reporting and public data sharing, I anticipated the need to contact authors and—in some cases—recalculate statistics from raw data. Therefore, I used the number of citations as a proxy for the impact of the papers, and focus this meta-analysis to papers with high number of citations. The papers from CHI 2020 are excluded because there is no information about citation count. It is important to consider that older papers are likely to have more citations than new ones [34], thus, I added the citations per year as a second criterion. The papers were ranked by citations per year and overall citation number, taking the absolute difference between those two values into account:

$$overall\_rank = citations_{rank} + citationsPerYear_{rank} - |citations_{rank} - citationsPerYear_{rank}| \quad (1)$$

The top 30 papers are included in the meta-analysis based on resource availability. This limitation is discussed in Sect. 5.

During the screening, papers that did not use human participants (2) or did not have a controlled experiment (1) were excluded. Four papers were excluded as they did not measure any typing performance (e.g., stress levels measured by a pressure-sensitive keyboard). The following 21 papers are included in the meta-analysis (ordered by year): 2009 [24], 2010 [1, 20], 2011 [13], 2012 [8, 12, 14], 2013 [33], 2014 [17, 30, 43], 2015 [25, 35, 38], 2016 [11, 15], 2017 [32, 46, 47], 2018 [22, 48].

2.2 Data collection

I manually collected the means and standard deviations from the papers they were fully reported in. For the papers with missing statistics or data, I contacted 14 authors via email. Five authors have provided necessary statistics, one author was unable to provide the necessary data, and eight authors did not respond within a month. Out of eight papers without response, seven were included partially and one was not possible to include.

2.3 Effect size and its variants

A simple effect size is a difference between two means. However, simple effect sizes from two experiments can be compared only if these experiments operationalized their dependent variable in the same exact way. For example, typing speed could not be compared to typing error rate. To enable comparison across experiments the effect size needs to be standardized . A standardized effect size can be calculated by dividing a simple effect size by a standardizer, which depends on the experiment setting [7]. One widely used standardized effect size is Cohen's d. However, Cohen's d is biased upwards when the sample size is smaller than 20 [37]. In this review, the average number of participants is 13.88 ($SD = 5.87$). To correct for this bias, I converted Cohen's d to Hedges's g.

For between-subjects designs, Hedges's g_s is used. For within-subjects designs, two types of Hedges's g are used. First, Hedges's g_{av} is advised for usage in meta-analyses by Lakens [23]. Second, Hedges's g_{rm} is a more conservative estimation for repeated-measures designs [23]. The latter can provide a safer estimation for experiment planning because the smaller is the planned effect size, the higher is the number of participants. Thus, the lower is the risk of getting insignificant results if the effect is present in the population. The formulas and the effect sizes calculation process are provided in the supplementary materials (see footnote 2).

[2]The analysis code, supplementary materials, and the data are available at https://osf.io/nqzpr

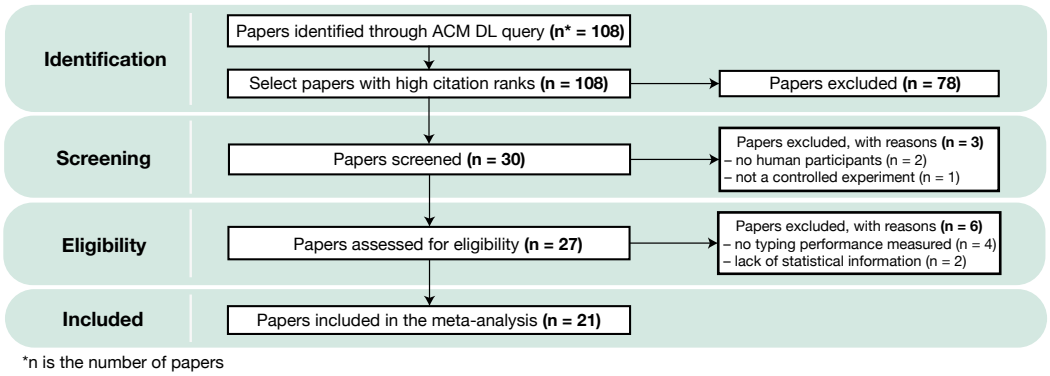

Figure 1: PRISMA diagram showing the process of papers selection in this analysis.

2.4 Data analysis

I included all effect sizes regardless of their statistical significance to avoid dichotomous decision based on p-values. To group the effect sizes as small, medium, and large, I followed the methodology used in the software engineering review [21]. I took 33% of the smallest effect sizes as small ones, 33% of the largest as large ones, and 34% in the middle are medium ones. For each of the three reference effect sizes, in addition to a pooled weighted estimate, I calculated the simple means, medians, and geometric means. For the geometric mean, zero values were ignored in the calculation. I used the exponential of the arithmetic mean of logarithms to express the geometric mean.

For each of the bins, I used the `metafor` [39] package for R to perform the multilevel meta-analysis to find the weighted average without making statements about the true population effect size. Thus, the model is used only to find these weighted averages and confidence intervals.

Overall, I included 21 papers with 26 studies in the meta-analysis where one study is an experiment of a paper with unique participants. If the participants are the same, the experiments are counted as one study.

3 RESULTS

First, I present the analysis results—the effect size ruler—for all included typing metrics. Then, I present the results for WPM (words per minute) separately because it was measured in 23 studies out of 26 and can be considered the most used typing performance metric. Since there were only 29 between-subjects effect sizes out of 475, separate results for them are not presented. Instead, the results of Hedges's g_{rm} and Hedges's g_{av} include Hedges's g_s in them.

The summary of the results is presented in Table 1 and Table 2. The distributions of the effect sizes are presented in Figure 2.

3.1 Results for all typing metrics

For small effect size, the mean for g_{rm} is 0.0965 ($SD = 0.0636$) is close to median 0.0933, and pooled weighted estimate 0.0955 with the 95% confidence interval of [0.0622, 0.1288]. The geometric mean (0.0754) is smaller than the mentioned above estimations. g_{av} mean is 0.1040 ($SD = 0.0701$) is also close to the median of 0.0962, and pooled weighted estimate of 0.1021 [0.0688, 0.1354]. The geometric mean is 0.1041, which is again lower than other estimations.

For medium effect size, the simple mean for g_{rm} is 0.3692 ($SD = 0.1044$) is close to all median 0.3638, geometric mean 0.3542, and pooled weighted estimate 0.3614 [0.3268, 0.3961]. The same goes for g_{av}: mean is 0.4122 ($SD = 0.1165$), median 0.4117, geometric mean is 0.3956, and pooled weighted estimate is 0.4039 [0.3685, 0.4392]. However, geometric mean is lower than other measures for both effect size types.

For large effect sizes, the difference between the estimations is the largest. g_{rm} mean is 1.4041 ($SD = 1.2748$), median 1.0068, geometric mean 1.1518, and pooled weighted estimate is 1.0414 [0.0414, 1.1891]. Similarly, g_{av} has the mean of 1.6186 ($SD = 1.3499$), the median of 1.0964, the geometric mean of 1.3344, and the pooled weighted estimate of 1.2378 [1.0702, 1.4054].

3.2 Results for WPM (words per minute)

The results of an additional analysis for words per minute are presented in Table 2. The results for WPM are larger than for all metrics. Pooled weighted estimates in Hedges's g_{rm} for small, medium, and large are 0.1270 [0.0676, 0.1863], 0.5090 [0.4389, 0.5792], and 1.6033 [1.3303, 1.8764] correspondingly. Similarly, Hedges's g_{av} estimates are 0.1366 [0.0773, 0.1959], 0.5471 [0.4754, 0.6187], and 1.9704 [1.6417, 2.2991]. As with all metrics results, all g_{av} effect sizes are larger than g_{rm}. The highest difference between the Hedges's g_{av} and Hedges's g_{rm} is visible for large effect sizes Figure 2 (C) and Figure 2 (D).

4 DISCUSSION

4.1 Comparison to other fields

In this section, I focus the discussion around Hedges's g_{rm} because it is the effect size I suggest to use for *a priori* power analysis. Table 3 shows the comparison of related reference effect sizes to the effect size ruler for typing experiments.

As in [21], Hedges's g are directly compared to Cohen's d from other studies. This comparison is valid because Hedges's g_{rm} is simply Cohen's d corrected for overestimation [26]. Hence, g_{rm} values are slightly smaller than Cohen's d values for small sample

Table 1: The results of the analysis of all typing performance metrics. Standard deviations for means and 95% confidence intervals for pooled weighted estimates are shown in the brackets.

Effect size	Type	Mean	Geometric mean	Median	Min	Max	Pooled weighted estimate
Small	g_{rm}	0.0965 (0.0636)	0.0754	0.0933	0.0000	0.1940	0.0955 [0.0622, 0.1288]
	g_{av}	0.1040 (0.0701)	0.0801	0.0962	0.0000	0.2289	0.1021 [0.0688, 0.1354]
Medium	g_{rm}	0.3692 (0.1044)	0.3542	0.3638	0.1977	0.5758	0.3614 [0.3268, 0.3961]
	g_{av}	0.4122 (0.1165)	0.3956	0.4117	0.2297	0.6496	0.4039 [0.3685, 0.4392]
Large	g_{rm}	1.4041 (1.2748)	1.1518	1.0068	0.5832	12.4194	1.0414 [0.8936, 1.1891]
	g_{av}	1.6186 (1.3499)	1.3344	1.0964	0.6515	12.4193	1.2378 [1.0702, 1.4054]

Table 2: The results of the analysis of WPM (words per minute). Standard deviations for means and 95% confidence intervals for pooled weighted estimates are shown in the brackets.

Effect size	Type	Mean	Geometric mean	Median	Min	Max	Pooled weighted estimate
Small	g_{rm}	0.1302 (0.0877)	0.0976	0.1329	0.0000	0.2850	0.1270 [0.0676, 0.1863]
	g_{av}	0.1395 (0.0943)	0.1041	0.1373	0.0000	0.3092	0.1366 [0.0773, 0.1959]
Medium	g_{rm}	0.5564 (0.1923)	0.5245	0.5155	0.2908	0.9527	0.5090 [0.4389, 0.5792]
	g_{av}	0.5960 (0.2055)	0.5617	0.5636	0.3112	1.0319	0.5471 [0.4754, 0.6187]
Large	g_{rm}	1.9727 (0.7508)	1.8400	1.7952	1.0475	3.6732	1.6033 [1.3303, 1.8764]
	g_{av}	2.3486 (0.9074)	2.1763	2.2651	1.0700	4.3729	1.9704 [1.6417, 2.2991]

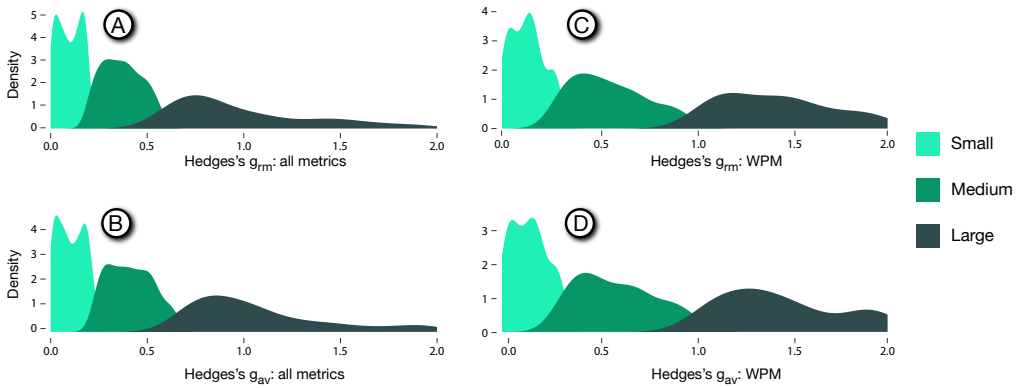

Figure 2: Distributions of the effect sizes: (A): Hedges's g_{rm} for all analyzed typing metrics; (B): Hedges's g_{av} for all analyzed typing metrics; (C): Hedges's g_{rm} for WPM (words per minute); (D): Hedges's g_{av} for WPM. The x-axis is truncated to show the distributions' shapes.

sizes. Nevertheless, the unit does not change: both Cohen's d values and Hedges's g_{rm} values are interpreted as a z-score in standard deviation units [37, p. 45]. HCI typing experiments estimate for small (0.10) and medium (0.36) effect sizes are lower than in other fields. This difference is due to the inclusion of all the comparisons from each study, including those with p-value > 0.05, and the actual difference between typing and other fields.

For large effect size, typing experiments estimated effect size of 1.04 which is larger than Cohen's convention (0.8), Psychology, Education, and Behavioral Sciences (0.9), but lower than the software engineering (1.40). Using pooled weighted estimated already addresses the extreme values in the large effect sizes bin. There were only ten studies which have found effect sizes of Hedges's g_{rm} larger than three and only one with an effect size of 12. The latter study has only five participants.

4.2 How to use the ruler?

This effect size ruler can be used in the following two ways:

(1) Researchers can use the ruler for *a priori* power analysis when no prior studies with similar measures and designs exist. To plan the experiment, the researchers can plan with

Table 3: The comparison of reference effect sizes in other fields to the typing experiment effect size ruler (See discussion about the effect sizes types in section 2.3 Effect size and its variants.)

Source	Small	Medium	Large
Cohen's convention based on human heights and intelligence quotients [5]	0.20	0.50	0.80
HCI typing experiments based on pooled weighted estimates with 95% confidence intervals (this work)	**0.10 [0.06, 0.13]**	**0.36 [0.33, 0.40]**	**1.04 [0.89, 1.19]**
Software engineering based on the median points [21]	0.17	0.60	1.40
Psychology, education, and behavioural sciences [26] based on the middle points according to [21]	0.15	0.45	0.90

all three reference effect sizes and compare a trade-off between increasing the number of participants and the gained power. They can further refine their plan by using the upper or lower bound of the interval estimates.

(2) Researchers can compare the newly found effect sizes to the ruler to identify whether it should be considered small, medium, or large.

Using the ruler for planning experiments can help researchers to avoid false-negative results caused by the lack of statistical power [7], or spending more resources on extra participants. Several tools allow doing *a priori* power analysis. For example, there is the R package `pwr` [2], software *G*Power* [10], and interactive web-based tool *Touchstone2* [9]. The knowledge of the small, medium and large effect sizes allows the researchers to enter all of them using these tools and assess the design trade-offs [42]. Using all three reference effect sizes and then assessing whether the sample size is reasonable can be a way to do *a priori* power analysis. For the α and $1 - \beta$ researchers can use 0.05 and 0.2 correspondingly [5, p. 56]. The tools can also provide the relevant effect sizes rulers as *Touchstone2* [9] provides the conventional Cohen's values already. Researchers can use Hedges's g_{rm} in the Cohen's f field to have safer planning with a higher number of participants but a lower risk of getting insignificant results.

The second purpose is to allow scientists to compare their found effect sizes to the reference small, medium, and large. It will enable future researchers to contextualize the magnitude of the improvement their new interaction technique has added. Both Hedges's g_{rm} and Hedges's g_{av} can be used while keeping in mind that g_{rm} is more conservative than g_{av}.

4.3 Reporting practices, data availability, and email availability

Out of 23 candidate papers, only nine of them reported all the necessary means and standard deviations for the meta-analysis. Out of the remaining 14 papers, only six authors answered emails about requests for the data, and five of those provided the needed information. From eight missing responses, seven papers reported partial data, which are included in this meta-analysis. In a survey [41], the authors have found that some researchers are not willing to share the data and information. It could be the case that some of the contacted authors had a similar opinion.

As of 2020, the papers included in this work are, on average, 5.95 (SD = 2.72) years old. I have manually searched for all the authors' email addresses on the ACM Digital Library, Google Scholar, and personal websites during this survey. While I have managed to send most of the emails, one of the authors in one paper delivery has failed. However, other authors in the same paper emails went through.

In the field of Psychology, Wicherts et al. [44] wrote 400 emails with data requests. Only 27% (n=108) of the contacted authors have sent the requested data, 16% (n=64) of which after the reminders. In [40], the authors found that finding a working email address drops by 7% with each year. The response rate in my work is comparatively high by receiving answers to 43% of my emails. This result could be due to manual search of the working emails, writing to each of the authors, and having the supervising professor's permission to include him in CC to increase trustworthiness.

One author replied that they were not able to provide sufficient information due to lack of access to the data. This issue has again been raised by [40], who surveyed 516 studies for data availability. They found that with every year, the availability of the data declines by 17%.

As a result of the above-mentioned issues, the situation constitutes a vicious circle. Lack of meta-analyses leads to authors not having a reference effect size. Without it, *a priori* power analysis is difficult to perform. Without a power analysis, researchers risk getting insignificant results. Due to *publication bias*, these papers are less likely to get published, and thus there is a file drawer problem raised [7]. If the papers do not get published, it becomes harder to do the meta-analyses, and the circle continues. This work aims to address the vicious circle by breaking it at the "No cumulative science" point.

As a recommendation for researchers in HCI field, I suggest to make the code and the data available on services such as osf.io to ensure the data availability, to plan the sample size using, for example, *a priori* power analysis, and to report full statistical results of the experiments.

5 LIMITATIONS

I acknowledge that my work has the following three limitations. First, since there is no common practice of preregistration in the HCI community, it is almost impossible to analyze the results of

unpublished studies [4, 41]. Therefore, I could not perform the publication bias assessment.

Second, the meta-analysis only includes articles published at CHI and the number of papers is relatively small. Future work should include more venues and also include more papers overall. However, even with 26 studies, the ruler already provides a more precise estimations than Cohen's conventions.

Third, this ruler can only be applied to typing experiments. Researchers doing experiments in different fields are advised to analyze each subfield separately. The code for the analysis and the formulas used for calculations are provided in the supplementary materials (see footnote 2).

6 CONCLUSION

In the HCI field, there is no local standard of the effect sizes. I contribute the effect size ruler for the typing experiments based on the meta-analysis of the literature published in CHI. Researchers can use this ruler to plan future experiments sample size using *a priori* power analysis and assess the magnitude of the found effect. This ruler can inspire other researchers to conduct systematic studies in their subfields, and aims to bring better science in the future.

ACKNOWLEDGMENTS

I am deeply indebted to Alexander Eiselmayer and Dr. Prof. Chat Wacharamanotham. This work would not have been possible without having them as my mentors. Their advice, help, unparalleled support, and constructive criticism have guided me throughout this project from the beginning to the very end. I am also grateful to authors who included the data in their projects, especially to the authors I was in contact with.

REFERENCES

[1] Xiaojun Bi, Barton A. Smith, and Shumin Zhai. 2010. Quasi-Qwerty Soft Keyboard Optimization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 283–286. https://doi.org/10.1145/1753326.1753367

[2] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. [n.d.]. Package 'pwr'. ([n. d.]).

[3] Mike W-L Cheung. 2014. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods* 19, 2 (2014), 211.

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

[5] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. (2 ed.). Taylor & Francis Ltd.

[6] Geoff Cumming. 2013. Cohen's d needs to be readily interpretable: Comment on Shieh (2013). *Behavior Research Methods* 45, 4 (2013), 968–971.

[7] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.

[8] Mark Dunlop and John Levine. 2012. Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2669–2678. https://doi.org/10.1145/2207676.2208659

[9] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300447

[10] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[11] Anna Maria Feit, Daryl Weir, and Antti Oulasvirta. 2016. How We Type: Movement Strategies and Performance in Everyday Typing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4262–4273. https://doi.org/10.1145/2858036.2858233

[12] Leah Findlater and Jacob Wobbrock. 2012. Personalized Input: Improving Ten-Finger Touchscreen Typing through Automatic Adaptation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2207676.2208520

[13] Leah Findlater, Jacob O. Wobbrock, and Daniel Wigdor. 2011. Typing on Flat Glass: Examining Ten-Finger Expert Typing Patterns on Touch Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2453–2462. https://doi.org/10.1145/1978942.1979301

[14] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2687–2696. https://doi.org/10.1145/2207676.2208662

[15] Aakar Gupta and Ravin Balakrishnan. 2016. DualKey: Miniature Screen Text Entry via Finger Identification. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 59–70. https://doi.org/10.1145/2858036.2858052

[16] M. Harrer, P. Cuijpers, T.A. Furukawa, and D.D. Ebert. 2019. Doing meta-analysis in R: A hands-on guide. *PROTECT Lab Erlangen* (2019). https://doi.org/10.5281/zenodo.2551803

[17] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed Endurance: A Metric to Quantify Arm Fatigue of Mid-Air Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1063–1072. https://doi.org/10.1145/2556288.2557130

[18] Kasper Hornbæk and Effie Lai-Chong Law. 2007. Meta-Analysis of Correlations among Usability Measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 617–626. https://doi.org/10.1145/1240624.1240722

[19] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[20] Eleanor Jones, Jason Alexander, Andreas Andreou, Pourang Irani, and Sriram Subramanian. 2010. GesText: Accelerometer-Based Gestural Text-Entry Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2173–2182. https://doi.org/10.1145/1753326.1753655

[21] Vigdis By Kampenes, Tore Dybå, Jo E Hannay, and Dag IK Sjøberg. 2007. A systematic review of effect size in software engineering experiments. *Information and Software Technology* 49, 11-12 (2007), 1073–1086.

[22] Pascal Knierim, Valentin Schwind, Anna Maria Feit, Florian Nieuwenhuizen, and Niels Henze. 2018. Physical Keyboards in Virtual Reality: Analysis of Typing Performance and Effects of Avatar Hands. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3173919

[23] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[24] Seungyon Lee and Shumin Zhai. 2009. The Performance of Touch Screen Soft Buttons. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 309–318. https://doi.org/10.1145/1518701.1518750

[25] Luis A. Leiva, Alireza Sahami, Alejandro Catala, Niels Henze, and Albrecht Schmidt. 2015. Text Entry on Tiny QWERTY Soft Keyboards. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 669–678. https://doi.org/10.1145/2702123.2702388

[26] Mark W Lipsey and Leona S Aiken. 1990. *Design sensitivity: Statistical power for experimental research*. Vol. 19. sage.

[27] 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.

[28] Tengteng Ma, Hasti Sharifi, and Debaleena Chattopadhyay. 2019. Virtual Humans in Health-Related Interventions: A Meta-Analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312853

[29] I. Scott MacKenzie and Kumiko Tanaka-Ishii. 2007. *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[30] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. 2014. Vulture: A Mid-Air Word-Gesture Keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1073–1082. https://doi.org/10.1145/2556288.2556964

[31] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med* 6, 7 (2009).

[32] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2558–2570. https://doi.org/10.1145/3025453.3025517

[33] 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* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2799–2802. https://doi.org/10.1145/2470654.2481387

[34] Henning Pohl and Aske Mottelson. 2019. How We Guide, Write, and Cite at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310429

[35] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/2702123.2702597

[36] Elizabeth Stowell, Mercedes C. Lyson, Herman Saksono, Reneé C. Wurth, Holly Jimison, Misha Pavel, and Andrea G. Parker. 2018. Designing and Evaluating MHealth Interventions for Vulnerable Populations: A Systematic Review. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3173574.3173589

[37] III Turner, Herbert M and Robert M Bernard. 2006. Calculating and synthesizing effect sizes. *Contemporary issues in communication science and disorders* 33, Spring (2006), 42–55.

[38] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry Using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 659–668. https://doi.org/10.1145/2702123.2702135

[39] Wolfgang Viechtbauer. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36, 3 (2010), 1–48. https://www.jstatsoft.org/v36/i03/

[40] Timothy H. Vines, Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24, 1 (2014), 94 – 97. https://doi.org/10.1016/j.cub.2013.11.014

[41] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448

[42] X. Wang, A. Eiselmayer, W. E. Mackay, K. Hornbaek, and C. Wacharamanotham. 2021. Argus: Interactive a priori Power Analysis. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 432–442. https://doi.org/10.1109/TVCG.2020.3028894

[43] Daryl Weir, Henning Pohl, Simon Rogers, Keith Vertanen, and Per Ola Kristensson. 2014. Uncertain Text Entry on Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2307–2316. https://doi.org/10.1145/2556288.2557412

[44] Jelte M Wicherts, Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. The poor availability of psychological research data for reanalysis. *American psychologist* 61, 7 (2006), 726.

[45] Nick Yee, Jeremy N Bailenson, and Kathryn Rickertsen. 2007. A Meta-Analysis of the Impact of the Inclusion and Realism of Human-like Faces on User Experiences in Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1240624.1240626

[46] Xin Yi, Chun Yu, Weijie Xu, Xiaojun Bi, and Yuanchun Shi. 2017. COMPASS: Rotational Keyboard on Non-Touch Smartwatches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 705–715. https://doi.org/10.1145/3025453.3025454

[47] Chun Yu, Yizheng Gu, Zhican Yang, Xin Yi, Hengliang Luo, and Yuanchun Shi. 2017. Tap, Dwell or Gesture? Exploring Head-Based Text Entry Techniques for HMDs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4479–4488. https://doi.org/10.1145/3025453.3025964

[48] Suwen Zhu, Tianyao Luo, Xiaojun Bi, and Shumin Zhai. 2018. Typing on an Invisible Keyboard. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174013