Local Standards for Sample Size at CHI

Kelly Caine
Clemson University
McAdams Hall
Clemson, SC 29634
caine@clemson.edu

ABSTRACT
We describe the primary ways researchers can determine the size of a sample of research participants, present the benefits and drawbacks of each of those methods, and focus on improving one method that could be useful to the CHI community: local standards. To determine local standards for sample size within the CHI community, we conducted an analysis of all manuscripts published at CHI2014. We find that sample size for manuscripts published at CHI ranges from 1 – 916,000 and the most common sample size is 12. We also find that sample size differs based on factors such as study setting and type of methodology employed. The outcome of this paper is an overview of the various ways sample size may be determined and an analysis of local standards for sample size within the CHI community. These contributions may be useful to researchers planning studies and reviewers evaluating the validity of results.

Author Keywords
Methodology; research methods; sample size; number of participants; N; evaluation; meta-HCI.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. H.1.2. Information Systems: User/Machine Systems - *Human Factors*.

General Terms
Human Factors; Design; Measurement.

INTRODUCTION
The CHI community is home to researchers from a wide range of disciplines including computer science, cognitive psychology, design, social science, human factors, artificial intelligence, graphics, visualization and multi-media design. Each of these disciplines has its own research method and manuscript preparation traditions. The CHI community also maintains strong connections with industry practitioners. In industry, methods traditions are often tempered with the need for pragmatism and efficiency.

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858498

Methodological rigor and appropriateness are key elements of the peer review process across disciplines. Within the CHI community, the instructions to reviewers about the importance of methodological validity are clear. Reviewers are asked to "assess the validity of the results" presented as a key element of any review. Indeed, the "Ensuring Results are Valid" section of the "Guide to a Successful Archival Submission" states, "reviewers often cite problems with validity as the reason to reject a submission" [1]. Therefore, understanding what constitutes a valid methodology is critical for authors who wish to have their research results accepted and published.

While there are many factors that contribute to the validity of a study such as how well the measures used represent the concepts of interest and how well the sample represents a population, the focus of this paper is on determining the sample size, which is often referred to as *N*.

The goal of the paper is to help readers understand the ways that sample size may be determined, understand the benefits and drawbacks of each method, and to create transparency about local standards for sample size within the CHI community.

Despite its importance to the validity of a study, determining the answer to the question, "how many users do I need?" is often not straightforward for researchers. For example, in an analysis of 55 empirical articles, researchers apologized for the size of their sample in 20%, indicating that even after a peer reviewed publication process, sample size questions remain [3]. There are many methods of determining the appropriate sample size for a given study, each with advantages and disadvantages (see Methods of Determining Sample Size). Reviewers often use sample size as a key determinant in determining validity of results.

Reviewers Incorrectly Use Sample Size to Reject Papers
Given the disciplinary breadth and crossover with industry practice, it is not surprising that there are so many methods for determining the appropriate sample size. It is also not surprising that reviewers often question the validity of the results reported in a manuscript based on the reported sample size, and subsequently recommend rejecting a submission based on a feeling that a "sample size is too small". This rationale, which we refer to as the *sample size fallacy,* has not been empirically studied in the CHI community (though it has been described; see [18]). However, the sample size fallacy is common in other fields,

such as the medical field [5; 6]). The criticism by reviewers that a sample size is insufficient has been demonstrated to be a "cover when reviewers cannot pinpoint, or are unwilling to admit, the real reasons why they dislike a proposal" or manuscript [6]. While there are valid reasons to reject a manuscript on sample size grounds, using sample size as a reason to reject a paper when a reviewer is unable to articulate and justify their real reasons is harmful for all parties involved: the authors, the PC members, the community and even the reviewer him or herself.

Why Are We Susceptible to the Sample Size Fallacy?

One reason the sample size fallacy is common is the threshold myth [6]. The threshold myth is that there is a threshold at which a sample size becomes "enough" to be valid. In reality, while the size of a sample is relative to the value of its findings, the relationship is curvilinear rather than square wave shaped (see Figure 1). That is, there is no meaningful cut-off point at which a sample size becomes "too small", inadequate or invalid [6]. Rather, the relationship between the value of a study and the size of the sample incrementally increases with each additional participant up to an asymptote, at which point there are diminishing returns for each additional participant. Therefore, while sample size can be justifiably criticized as inadequate to determine whether there is a reliable effect at a certain level of confidence, there is no point at which a sample size can be justifiably criticized as "too small" without qualification.

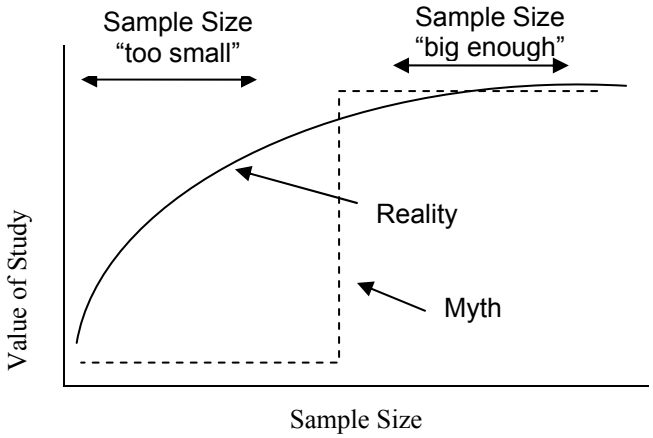

Figure 1. The Sample Size Threshold Myth (adapted from Figure 1, [6])

Despite this, for all studies that involve research participants or "users", researchers must still decide a specific point at which they will cut off or stop testing, observing or interviewing participants. Thus, researchers and reviewers need a way to assess the validity of the size of the sample for a given manuscript or proposal. How is this currently accomplished?

METHODS OF DETERMINING SAMPLE SIZE

There are many methods for determining how many participants are required for a research study including power analysis, saturation, cost or return on investment (ROI) analysis and guidelines, including local standards. We describe these methods below and discuss the limitations and criticisms of each.

Prospective Power Analysis

For quantitative studies, the formal, statistically defensible method to determine how many participants you need for a study where you will draw statistical inferences is a prospective power analysis [15]. A power analysis gives you the probability of rejecting a null hypothesis given the alternative hypothesis is true. Both post-hoc and a priori power analysis consider the type of statistical test you plan to conduct, the significance level or confidence you desire, the size of the effect you expect, the level of "noise" in your data, and the sample size (N). Because these four factors interact with one another, if you know three, you can determine the fourth. This means you can use this type of analysis to determine the sample size you need to be able to confidently and reliably detect an effect as long as you know, or can estimate, the significance level, noise level and effect size [23].

Limitations and criticisms

While power analysis is recognized as a rigorous and defensible method of determining sample size, it is not without limitations or detractors (e.g., [7]). One significant limitation for researchers, who are interested in innovative technology, is that power analysis requires existing quantitative data about the research topic. To conduct a power analysis a researcher must know the significance level, noise level and effect size in order to compute the necessary sample size. For researchers working with new technologies, as is common among the CHI community, this type of preliminary data often does not exist (though one recommendation is to use a general estimate, for example, Cohen's $d = 0.5$, alpha $= 0.05$, and $\beta = 0.85$).

Beyond this, power analysis itself is built upon statistical heuristics. For example, the values for small, medium, and large effect sizes are themselves guidelines produced by an expert, Jacob Cohen [14]. Even the "p-value" was meant by Ronald Fisher, its inventor, to be, "an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look" [26].

Saturation

Data saturation is the point during qualitative data collection at which no new relevant information emerges [16]. Because saturation is not known until it is reached, it is not possible to determine in advance the sample size that will be required before the amount of new information per participant tapers off

Limitations and criticisms

Saturation can be difficult to justify to both stakeholders and reviewers because 1) it cannot be predicted in advance and 2) providing evidence that saturation has been reached is difficult. Furthermore, for the researcher, saturation makes study planning difficult because a researcher does

not know in advance when saturation will be reached. This can lead to recruiting too many or too few participants and difficulties coordinating data collection.

Cost and Feasibility Analysis

There are two methods to determine sample size based on resource limitations: cost and feasibility analyses.

Cost or ROI Analysis

When a researcher already knows the limit of funding they have available for research, a cost analysis can help them determine how many participants they can recruit. The simplest, least informed analysis is:

$$\frac{(total\ \$\ available\ to\ pay\ participants)}{(\$\ per\ participant)} = number\ of\ participants$$

Other study variables such as reducing the duration of the study or changing the number of conditions in an experiment, for example, can be manipulated to increase or decrease the number of participants possible.

Feasibility Analysis

Besides monetary costs, there are often other constraints a researcher is aware of as they plan a study. Constraints include: time available to complete a study, participant availability, number of participants that exist (e.g., the population of astronauts is much more limited than the population of laptop users [27]; the availability of surgeons to participate in studies is much less than nurses), number of prototypes, number of researchers available, and space.

With a feasibility analysis, a researcher can use the constraints they know to guide the number of participants they sample. For example, if a researcher only has four prototype devices, a study that takes three hours to complete, and a one-week window within which participants are available, it will not be possible to test hundreds of participants. Instead, the researcher could use these constraints to determine the maximum sample size they can feasibly test given this situation. When a feasibility analysis is used, the typical recommendation is that the researcher should report both the sample size recommended by a power analysis, for example, and the size used for the study, along with an explanation of the constraints that led to the smaller sample size.

Limitations and criticisms

While all study planning involves feasibility analysis even if researchers may not like to admit it, constraints should ideally be only a part of what helps a researcher choose a sample size. Indeed, up to a point, there is only a benefit to each additional participant in a sample size when considered and weighed against the cost and feasibility of conducting the study. Furthermore, despite their ubiquity, cost and feasibility are rarely mentioned in manuscripts reporting research results. This type of analysis is more typical in industry than academia, though some argue it should be more heavily relied upon in scientific funding decisions (e.g., [8; 17]) precisely because it considers the trade-off between cost and return on investment.

Guidelines

There are two types of guidelines for determining sample size: recommendations by experts and local standards.

Recommendations by experts

Experts are people who have worked in a field with particular success and often for a long time. Because of this experience, they are viewed as trusted sources of valuable information about a topic. For example, we conducted a literature search for information about sample size with respect to usability studies and found highly expert recommendations ranging from 4 ± 1 (for think aloud studies [25]) to 10 ± 2 [21] with others recommending a grounded procedure which starts with an estimate, observes the data collected for the estimate and then reevaluates [11]. See [10, pg. 108] for a summary of expert recommendations about sample sizes for various methods.

Limitations and criticisms of recommendations by experts

While expert recommendations are available for some methods (e.g., usability studies), it is difficult to find recommendations for other types of research methods. For example, we found very little in the way of expert sample size recommendations for surveys. Furthermore, relying expertise shares many limitations with local standards.

Local Standards

Local standards are guidelines based on similar or analogous studies that have already been published. Researchers can find out about the local standards in their organization or community by asking colleagues how many participants they have used for studies similar to the one being planned. If a researcher plans to publish his or her work, s/he can determine local norms by reviewing published papers from the venue of interest.

Limitations and criticisms of local standards

First, relying solely on prior work could lead researchers to make an *argumentum ad populum*, fallaciously concluding that simply because many others have used some sample size, it must be appropriate. Rather, researchers must realize that choosing an appropriate sample size depends on a number of factors such as the study approach, effect size and availability of participants. Second, for a researcher to be able to consult local standards, a number of conditions must be met. The researcher must be part of an organization that has other researchers or have a network of colleagues that the researcher feels comfortable querying about sample size practices. While many researchers work in a setting with these amenities, others do not and will thus have difficulty obtaining information about local standards.

The second recommendation for obtaining local standards is to consult recently published papers from the venue where a researcher plans to submit research findings for publication. However, recent summary information about sample size is not available (but see [9] for guidance), so getting this information in any other than an anecdotal way would require a great deal of individual time and effort.

Need for Study

To assist researchers with understanding local standards about sample size in the CHI community, there is a need for a systematic analysis of community-wide practices.

METHODS

We conducted a systematic literature review of all manuscripts published at CHI2014 and manually extracted data from each manuscript. We collected: the contribution type, presence or absence of a user study, sample size, number of studies per manuscript, setting, method, manuscript length, award status, student status and gender breakdown of participants. We used this data to generate summary information about typical sample size at CHI.

Manual Extraction of Category Data

We extracted methods data from each paper/note manually. First, we created a spreadsheet with one row for each manuscript. The spreadsheet contains a column for each of the following fields: title, authors, number of studies reported, method, setting, N, N_male, N_female, Age_data, Age_measure, students (yes, no), funded (yes, no), approach (qualitative, quantitative), other demographics (free text), justification (free text), and notes (free text).

Next, a research assistant used a web browser to view the CHI2014 Proceedings table of contents via the ACM Library. The research assistant then read the title and abstract of an individual manuscript to get an overview of the content. Next, the research assistant opened the pdf of the paper and sought the content required to fill in each field. The following instructions were provided:

1. Glance through the paper to get a sense of what's there; look for the "methods" section, if available.
2. Find the place in the paper that describes the methods of the study. This will vary by paper.
3. Search terms that may help you find the information you need: participants, subjects, male, female, women, men, demographics.
4. Copy and paste the information from the pdf of the paper into the correct column of the spreadsheet. Be sure to include the text you used to determine the method and the number of participants. Use the notes field to explain anything that does not fit neatly into a category.

Research assistants were instructed to use the authors' description/categorization of the work, rather than their own judgment. For example, if an author described their work as "ethnography", we counted that in the "ethnography" category, even if the author described conducting a focus group or interview as part of the ethnography. Similarly, if an author described a study as an "experiment" we counted that in the "experiment" category rather than "mixed-methods" even if data collected included a post-task survey or questionnaire. However, when authors reported their studies using a generic descriptor (e.g., "user study"), we assigned the work to the category that was the closest methodological match to the work they described. For example, if an author called a study a "user study" then went on to describe an experimental set-up with between-subjects tests, we categorized that study as "experiment".

Once the spreadsheet was complete, at least one additional researcher reviewed each entry for accuracy.

Analysis

In consultation with a professional statistician, we considered multiple analysis methods to deal with skewed data and to satisfy the constant variance assumption. We settled on the use of non-parametric tests (e.g., Mann-Whitney U, Kruskal–Wallis), where appropriate. Non-parametric tests were used to compare distributions of sample sizes among levels in categorical variables due to the right-skewness of the distribution of sample size. When the constant variance assumption was not satisfied, we used a natural log transformation of the sample size. A level of 0.05 was used for all tests of significance.

RESULTS

We present the results of our data analysis in the following sections: descriptive statistics, number of studies per manuscript, setting, approach, method, manuscript length, student status and gender breakdown.

Descriptive Statistics

In 2014, there were 465 manuscripts published at CHI [31]. This represents 13% of all manuscripts published CHI between 1994 and 2014 [24]. Of these, 423 included a study with research participants. For simplicity, we will call these *manuscripts with user studies*. The remaining 42 manuscripts were theoretical, methodological, critical, modeling or technical contributions (see Table 1).

Manuscripts	465	100%
User study	423	91%
No user study	42	9%

Table 1. Percent of manuscripts containing user studies.

Number of Studies per Manuscript

Sometimes multiple user studies were reported within one manuscript. These manuscripts came in two styles: "multiple studies" and "mixed-methods". For the purposes of analysis, we considered more than one study with *different* participants reported in a single manuscript "multiple studies" whereas studies using the *same participants* we (and in most cases the authors) labeled "mixed-methods". For example, if a manuscript reported a focus group and an interview with the same participants, we considered this a mixed-methods study. On the other hand, if a manuscript reported an interview and a focus group with different participants for each, we considered this multiple studies. We excluded author-identified pilot studies. A majority of manuscripts reported a single user study (see Table 2).

Number of Studies Reported	Manuscripts	%
Single	289	68
Multiple	134	32
2	*101*	*24*
3	*25*	*6*
4+	*8*	*2*

Table 2: Manuscripts reporting single vs. multiple studies.

Given that some manuscripts reported multiple studies, there were more user studies represented than the total number of manuscripts that reported a user study. There were 606 user studies reported at CHI2014. The mean number of studies reported per manuscript is 1.4. In most of the further analyses we use user study (N = 606), user studies that report sample size (N = 560) or user studies that report a sample size and are not extreme (N = 519; see Outlier Analysis) as the unit of analysis.

Sample Size Reporting

The vast majority of user studies included the number of participants who participated in the user study (see Table 3). However, forty-six studies were published without reporting the size of their sample. Of these, many (39%) were analyses of an existing corpus of data (e.g., blogs, twitter stream). We counted these as user studies since they reported data about users. However, most of these provided the number of data points (e.g., tweets) rather than the number of individual people who may have contributed data. A notable exception is the manuscript that reported the largest N, [19], which reported the number of twitter users, rather than tweets, which was thus counted as reporting an N. The remaining twenty-eight user studies omitted sample size. They were four ethnographies, eleven experiments, two field studies, one focus group, two observations, one participatory design, one usability test and six other user studies that did not fit into one of the defined methods categories (see Section "By Method").

User study	606	100%
Report N	560	92%
Do not report N	46	8%

Table 3. Percent of user studies that provide sample size.

Descriptive Data about Sample Size

For the 560 studies where the number of participants was reported, sample size ranged from 1 – 916,000 (see Table 4 and Figure 2).

Range	Mean	SD	Median	Mode
1 - 916,000	4,119	42,856	18	12

Table 4. Descriptive statistics of sample size at CHI2014.

The most commonly reported sample size was 12. Fifty-seven studies, a full 10% of all user studies, reported a sample size of 12. Twenty percent of studies reported a sample size of ten or less, half of studies reported a sample size of less than 18 and seventy percent of studies reported a sample size of less than 30.

Mean and median sample size differ widely indicating skewed data. A scatter plot that showed a long tail of large sample sizes (see Figure 2) confirmed this skew.

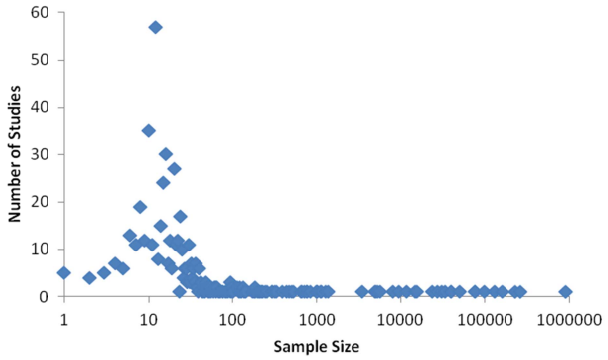

Figure 2. Scatter plot of sample size (log transformed).

Method Frequency

Manuscripts reported a variety of methods for user studies. The type of method employed in a user study was not evenly distributed; some methods (e.g., experiments) were much more frequently represented than others (e.g., diary study). The most common user study method was an experiment (41%) and the second most common was an interview (10%; see Table 5). We combined methods that did not fit within a category into an "other" category. This category included names such as, "first use study", "critical design", "user-elicitation" and "real world deployment".

Type of Method	Frequency	%
Data/corpus/log	30	5%
Diary study	4	1%
Ethnography	13	2%
Experience sampling	3	>1%
Experiment	249	41%
Eyetracking	14	2%
Field Study	32	5%
Focus Group	7	1%
Interview	60	10%
Mixed-method	28	5%
Observation	38	6%
Participatory design	34	6%
Survey/Questionnaire	8	1%
Usability test	21	3%
Other user study	65	11%
Total	606	100%

Table 5: Method Frequency

Outlier Analysis

Because data about sample size were skewed, prior to additional investigation, we conducted an outlier analysis using individual box plots for sample size by setting to identify extreme sample sizes. Based on this analysis we eliminated all cases that were more than three interquartile ranges away from the first or third quartile and the most extreme 5% that were more than 1.5 interquartile ranges away from the first or third quartile. There were 41 studies identified as outliers based on sample size by setting (in person vs. remote). We omitted these outliers from further analysis leaving 519 studies as the final dataset (see Table 6).

For in-person studies we eliminated five experiments, two eyetracking studies, one field study, one focus group, one interview, two mixed methods studies, one usability test and one "other" (14 in total).

For remote studies, we eliminated eight dataset/corpus studies, eleven experiments, one field study, one mixed-methods study, two observations, and one survey/questionnaire (24 total).

User studies that report sample size	560	100%
Included	519	93%
Excluded	41	7%

Table 6. Percent of user studies included in analysis.

Setting (In-person vs. Remote)

Seventy percent of user studies were conducted in-person, while 20% of studies were conducted remotely (e.g., interviews conducted via video chat; experiments via a website). Three percent of studies used a combination of in-person and remote methods, and four percent did not report whether the study was conducted in-person or remotely. There was a significant difference in the sample size for studies conducted in-person vs. those conducted remotely (ln(n) transformed Mann-Whitney $z=10.27$, $p < 0.001$). Studies conducted remotely had a much larger sample size than those conducted in-person (see Table 7). Because sample size varied drastically by setting, we report results for each setting (in-person vs. remote) separately.

Setting	**Studies**	**Mean**	**SD**	**Median**
In-person[1]	379	18	12	15
Remote[1]	105	197	285	77
Combo	14	15	6	15
Not reported	21	20	11	20
Total	519	54	147	16

[1]Note: ln(n) used in the analysis to compare in-person and remote sample size distributions.

Table 7: Setting (in-person vs. remote)

	In Person			Remote		
	N[1]	Mean	SD	N[1]	Mean	SD
Approach						
Qualitative	163	14	9	41	155	257
Quant.	216	20	12	64	224	300
Method*						
Data/corp	-	-	-	4	549	359
Diary	-	-	-	3	26	10
Exp. Samp.	-	-	-	3	351	562
Ethno.	5	6	4	-	-	-
Other	24	12	10	1	52	-
Usability	58	16	12	3	148	127
Interview	34	16	10	12	15	6
Part Des.	6	20	10	1	23	-
Experiment	182	20	12	31	224	272
-Within	117	17	9	8	252	278
-Between	46	26	16	16	236	325
-Mixed	15	25	13	6	188	87
Observ.	24	18	11	10	97	156
Field	14	19	16	12	89	215
Mixed	15	21	11	6	106	82
FG	5	21	7	-	-	-
ET	12	21	8	-	-	-
Survey	-	-	-	19	371	368
# of Studies						
Single	178	20	12	47	193	268
Multiple	201	16	10	58	200	301
Manuscript*						
Paper	316	18	12	92	208	299
Note	63	19	12	13	119	136
Award*						
Best	18	11	5	3	71	98
HM	62	19	11	16	191	246
None	299	18	12	86	203	297
Funding						
Funded	231	19	12	63	194	285
Not funded	148	17	11	42	201	289
Student*						
Students	81	22	15	10	51	57
Non	145	16	10	70	228	315
Gender	286			65		
Women		7	6		92	131
Men		10	6		117	183

*log(n) used in analyses to satisfy the constant variance assumption

[1] N here refers to the number of user studies; Mean and SD refer to sample size.

Table 8. Descriptive Data by Setting.

Approach: Qualitative vs. Quantitative

A qualitative approach favors data in the form of rich verbal description, while a quantitative approach favors numeric data [28]. Determining whether a study is qualitative vs. quantitative is multifaceted. Qualitative data, such as transcribed interview data, can be counted and analyzed quantitatively; in this case, the data are qualitative but the

analysis is quantitative. So what then distinguishes qualitative from quantitative? Is it the type of data collected? The theoretical approach? The analysis method? We chose the following criterion: presence of statistical analysis, in part based on the work of [12]. Those studies that included a statistical analysis we categorized as "quantitative". Those that did not include statistical analysis, we categorized as "qualitative". This definition is clearly imperfect (see the limitations section for a discussion of this), but it provides a useful analog to methods of determining sample size (i.e. a power analysis is applicable only when research involves statistical analysis).

Overall, we found that 44% of studies were qualitative and 56% were quantitative. [12] reported a similar breakdown in HCI studies: 50% of studies he analyzed reported statistical tests.

For both in-person and remote studies, we found a difference in sample size between qualitative and quantitative studies ($z=-5.82$, $p<0.001$ and $z=-2.23$, $p=0.026$ respectively using Mann-Whitney two-sample test). In both cases, qualitative studies had a lower mean sample size than quantitative studies (see Table 8).

By Method

As described above (see section on method frequency), manuscripts reported a variety of methods for user studies. Some methods, for example eye-tracking studies, were all conducted in-person. On the other hand, some interview studies were remote while others were conducted in-person or as a combination of in person and remote. In addition to testing differences across methods, we also specifically tested to determine whether there were differences in sample size across type of experiment (within, between and mixed designs).

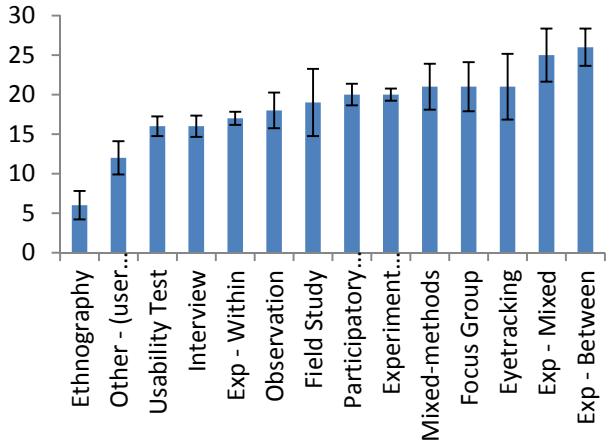

Figure 3. Mean sample size by method (SE) for in-person studies.

In-person

The sample size for in-person studies varied by the type of method (Kruskal-Wallis $X^2=33.67$, $p<0.001$; see Figure 3 and Table 8). Ethnography had the lowest mean sample size (6) while mixed methods and eye-tracking studies had the highest (21; see Table 8).

Experiment Type: For in-person studies, the sample size significantly varied by the type of experiment used (i.e., between-subjects, within-subjects, and mixed design; Kruskal-Wallis $X^2=10.57$, $p=0.005$). The average sample size for within-subjects experiments (17) was smaller than the average for between-subjects (26) or mixed designs (25).

Remote

The sample size for remote studies varied by the type of method used, Kruskal-Wallis $X^2=45.48$, $p<0.001$ (see Figure 4 and Table 8). Interviews had the lowest mean sample size (15) while dataset/corpus had the highest (549).

Experiment Type: For remote studies, sample size was similar across experiment type (between, within vs. mixed studies; Kruskal-Wallis $X^2=0.45$, $p=0.796$).

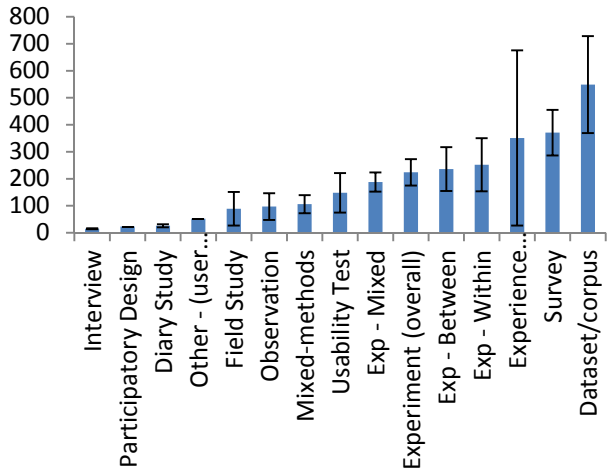

Figure 4. Mean sample size by method (SE) for remote studies.

Paper vs. Note

Manuscripts published at CHI fall into one of two length categories. Papers "must break new ground and provide complete and substantial support for its results and conclusions," whereas notes are "more focused and succinct" and are "likely to have a smaller—yet still significant—scope of contribution" [2]. Papers are at maximum ten pages, whereas notes are at maximum four.

	Single	Multi	Percent Multi-Study
Paper	233	117	33%
Note	58	15	21%

Table 9: Proportion of Manuscripts Reporting Multiple Studies by Type (Paper vs. Note)

Notes are not expected to cover the entire iterative design cycle and may instead focus on providing depth in a specific area. Therefore, we expected a difference in the percent of multiple study manuscripts to single study

manuscripts in papers vs. notes. Not surprisingly, more papers than notes reported multiple studies within one manuscript (Fischer exact test: p = 0.037; see Table 9).

Furthermore, because papers must "provide substantial support for its results" while notes are expected to be "more succinct", we expected that papers would report larger Ns.

Contrary to our expectation, there was no significant difference in the sample size of studies reported in papers vs. notes for in-person studies, Mann-Whitney z=0.34, p=0.73, or for remote studies, Mann-Whitney z=-0.43, p=0.67 (see Table 8).

Single and Multiple Studies Have Similar Sample Size

As described in the section on number of studies per manuscript, sometimes multiple user studies were reported within one manuscript (see Table 9).

In-person

For in-person studies, the sample size varied by whether a study was part of a manuscript that reported multiple studies vs. those that reported a single study (z=2.77, p=0.006). Studies part of single study manuscripts reported a higher sample size (20) than those that were part of a multi-study manuscript (16; see Table 8).

Remote

For remote studies, the sample size was similar for studies that were part of a manuscript that reported multiple studies and those that reported a single study, z=0.05, p=0.956 (Mann-Whitney; see Table 8).

Funding

Research presented at CHI is funded in a variety of ways including national and international funding agencies (e.g., NSF) and by industry. We examined the acknowledgments section of manuscripts to determine whether authors noted a funding source. We expected papers that were funded to report larger sample sizes than papers that were not funded.

For both in-person and remote studies, sample size was similar for funded and unfunded studies (Mann-Whitney z=-0.839, p=0.401 and z=-0.278, p=0.781; see Table 8).

Student Participants

Almost three quarters (71%, after removing studies that failed to report an N, outlier analysis, etc.) of the user studies reported whether the participants in the study were college students. Of these, 19% of studies reported college students as the sole participants.

In-person

For in-person studies, the sample size varied by whether participants were students, Mann-Whitney z=2.49, p=0.013 (see Table 8). In-person studies with students reported a higher sample size (22) than those with non-students (16).

Remote

For remote studies, the sample size significantly differed between student and non-student studies, Mann-Whitney z=-2.15, p=0.031 (see Table 8). The average for remote studies without students was higher (228) than the average for remote studies with students (51).

Gender

Almost three quarters of studies (71%; after removing studies that failed to report an N, outliers, etc.) of the user studies reported the gender breakdown of participants.

For both in-person and remote studies, we found a difference in the number of women vs. men that participated in user studies (Wilcoxon Signed Rank S=7851, p < 0.001 and S=327, p=0.012 respectively; see Table 8). In both cases, fewer women than men participated user studies (see Table 8).

DISCUSSION

The goal of this paper is to help readers understand the ways that sample size may be determined, the benefits and drawbacks of each method, and to provide transparency about local standards within the CHI community. An understanding of community practice can complement existing methods of sample size determination.

The Range of Sample Size at CHI is Large

One key takeaway from this analysis is that the range of sample size is extremely large (from N=1 to N=916,000). This is likely due to differences in setting (in person vs. remote), approach (qualitative vs. quantitative) and method choice (e.g., experiment vs. ethnography) that reflect the diversity of the CHI community. No one sample size fits all; researchers and reviewers must take into account a huge number of factors including the research question, method, and availability of participants when determining the appropriateness of sample size for a particular study.

Small N Studies are Publishable

Another key takeaway is that studies with a "small" sample size are publishable. Indeed, seventy percent of those studies published at CHI2014 reported a sample size of less than 30. Twelve was the most common sample size across studies accounting for a full 10% of all studies. The median sample size in 2014 (18) is in line with sample size reports from 1983 – 2006 (ranging from five to 29) [9] indicating sample size at CHI has remained consistent over time. In one respect, the finding that small N studies are published at CHI should not be surprising. The research presented at CHI is often the first research conducted in an area, and studies of new technologies must often start small, "sometimes even with an n of 1 because of cost and feasibility concerns" [7]. Furthermore, studies with a small sample size can reveal the most obvious usability problems [4]. Finally, from a return on investment perspective, small sample sizes "can produce more projected scientific value per dollar spent than larger sample sizes" [7]. On the other hand, especially for quantitative work, it appears that many studies published at CHI are underpowered even to find large effects: a two condition, within-subjects study with 17 participants (the mean for in-person, within-subjects experiments), has a power of 0.49 to find a large effect.

Small N Studies, Especially, Should be Replicated

The median and mode sample size findings, the mean sample size findings for in-person studies, and the finding that many quantitative studies are underpowered, add a new facet to Greenberg and Buxton [18] and Reed and Chi's [13; 29] recommendation that the CHI community embrace replication research. If many studies presented at CHI choose sample sizes because of cost and feasibility concerns, as suggested by [7], it is critical that these studies be replicated for the scientific integrity of our field.

Student Status, Gender and Age of Participants

Despite typically being "weird" (Western, educated, industrialized, rich, and from democratic societies [20]) and unlikely to be representative of even the typical student population [30], college students are often used as a convenience sample because of their proximity to researchers. About half of studies overall and three quarters after removal of outliers, etc., reported whether students were used as participants. Of these, 28% overall and 19% after outlier removal reported using students as the sole participants. For comparison, in 2006, 57% of manuscripts reported student participants [9]. One reason the percentage of studies reporting the use of student participants may be falling is because of the increasing incidence of remote studies; the importance of proximity of students to researchers diminishes for remote studies.

Only around half of overall studies and three quarters after removal of outliers, etc., reported the gender ratio of their sample. This remained roughly unchanged since 2006 [9]. More men than women participated in user studies, though the gap appears to have decreased since 2006 [9]. Notably, similar numbers of women and men participated in remote studies, though there was a statistical difference. Similar to our speculation around student participants, it may be that the increasing availability of remote study facilities can help balance the gender of study participants.

While we did not report it in the results section due to the limited availability and nature of the data and space constraints, it is worth noting that only 58% of manuscripts reported a measure of age of participants. Mean ages were between five and 81. While it is clear that participants ranged in age from children to older adults, it was difficult to ascertain the central tendency of the age of participants.

Summary: Recommendations for Authors

Based on this analysis, we have derived a number of recommendations for authors:

1. Use methods that are appropriate to your approach and analysis strategy to determine sample size. For example, if you plan to perform statistical analyses, use a power analysis; for qualitative work use saturation.
2. *Always* report sample size and the methods used to determine sample size.
3. Include all relevant demographic information (e.g., gender, student status, age) about the sample.
4. Include supplementary information such as power analysis and effect sizes (for quantitative studies, [22]) or the saturation criterion (for qualitative studies, [16]).
5. Note any constraints with respect to sample size. If cost or feasibility concerns played a part in sample size determination, note these [8; 17]. Explain how these limitations affect the interpretation of your findings.

Using Local Standards: Caution Required

Relying on local standards in isolation to assess the validity of a sample size should not be considered "best practice". Using a local standard (from this paper, or any other source), exclusive of other considerations, may lead to a sample size determination that is inappropriate for your research question; an inappropriate sample size could jeopardize the validity of your study. Best practice for choosing a sample size depends on a number of considerations including the disciplinary traditions of your approach, the type of analysis planned, the size of a population, and the cost and feasibility of the study.

Takeaways for Reviewers

Because reviewers and PC members are subject to the sample size fallacy (e.g., [6; 15]), they should, at a minimum, be reminded of this fallacy and asked to consider the method of determining sample size (e.g., saturation, expert recommendation) when assessing the size of a sample reported in a manuscript. If the HCI field is like other fields, then the review process, and the reviews provided to potential authors, would be improved if reviewers focused on unearthing the underlying criticisms of a paper, rather than claiming small sample size as a rejection "cover" [6]. Going beyond the minimum, for quantitative studies, we recommend that the CHI community instruct authors to conduct a power analysis prior to conducting a study, and to note this practice as part of their method section. We also recommend that reviewers insist on the inclusion of a power analysis and rely on this evidence about the adequacy of the sample size.

Another consideration for reviewers is that some reported sample sizes are likely "excessive/too big" which has the potential to lead to the publication of results that are not do not have a practical impact for users, even when statistically significant (i.e., Type I error).

The Evolution of Including User Studies in CHI Papers

In addition to these findings about sample size, a number of other notable findings with respect to general publishing practices at CHI emerged from these data.

First, 91% of manuscripts from 2014 reported a user study of some kind. For comparison, from 1983 to 2006 the portion of manuscripts that included an evaluation ranged from 50% to 97% [9]. This may indicate that reviewers and the PC increasingly expect CHI papers to include a user study. However, perhaps because some portion of the CHI community is not used to including a user study, 28 user studies failed to include the size of the sample. We found it

surprising that manuscripts that reported a user study were accepted and published without reporting a sample size, and suggest that any paper published at CHI that includes a user study should report the sample size.

Second, the most commonly reported type of user study is an experiment (41% of all user studies). The second most common is an interview (10%). Following these are many other methods that were less represented at CHI2014. It may be that researchers in the CHI community choose experiments most often because the research questions they are seeking to answer are best addressed through experiments. On the other hand, it could be that reviewers prefer experiments because they perceive them as more rigorous than other methods. While we cannot answer that question with these data, it is one we would be interested in exploring in the future.

Limitations

One major limitation of this study is that we only analyzed data from the proceedings of CHI for a single year, even though we know there is sample size variation across years for studies published at CHI [9]. We chose to analyze data from a single year for three reasons. First, we needed a way to limit our sample size. Previous work has sampled 358 papers [9] and 360 [32]; we used this range as a starting point, though we ended up sampling more manuscripts (465) so that we could cover an entire proceedings. For our purposes, we thought it was preferable to survey an entire proceedings, rather than randomly sampling multiple proceedings, because it provided not only an overview of sample size, but also of the CHI proceedings itself. For example, it allowed us to investigate the frequency of methods (e.g., we were surprised to learn that 41% of studies published at CHI2014 were experiments) and to compare sample size across variables such as method. Finally, for developing local standards, we thought recent data would be the most valuable.

Another limitation, although we feel a defensible, purposeful limitation, is the inclusion of only papers published at CHI. We fully acknowledge that our community is larger than the numbers represented by papers published at CHI. However, as articulated by [24], "considering the relevance of the CHI conference to the field of HCI, an analysis on the CHI articles should enable us to attain a fair overview of the field." Future work should consider papers published in venues that more fully represent the community (e.g., CSCW, UbiComp, TOCHI).

Another limitation of this study is that the data are based on self-report. That is, authors self-report the methods they use and the number of participants they study. It is possible that some of this information is inaccurate or that different communities would call methods by different names. For example, while some in the CHI community would call a study ethnography, some in the anthropology community may not. We addressed this by consistently accepting authors' own description of their methods. However, authors' descriptions were sometimes difficult to resolve within our, admittedly limited and imperfect, categorization scheme. For example, in some cases, authors called their studies "experiments", but no statistical analyses were reported. These cases led to somewhat confusing results such as studies being categorized as qualitative experiments. Furthermore, some authors did not report the size of their sample, gender breakdown, student status, etc. limiting our ability to analyze this data.

Yet another limitation of this work is that the extraction of data from papers was manual, rather than automatic. Humans are fallible and it is possible that we recorded some information incorrectly. We attempted to mitigate this threat by having two or more researchers review each set of data and agree about what was extracted.

Finally, caution is required when using local standards. For example, using solely local standards in quantitative work is considered ineffective by statisticians [14; 15]. Indeed, if the researchers who have published at CHI previously have chosen their sample size based on cost or other constraints analysis and we follow them, our sample size will be similarly constrained; local standards then may not be considered a "best practice", but rather a pragmatic norm.

Future Work

If local standards are considered by the community to be an important method of determining the number of participants for studies published at CHI, it would be beneficial to create a more systematic way of capturing sample size and related data from each paper published. This information could be requested at the time of submission or publication and published via the ACM or SIGCHI website.

CONCLUSION

The size of a sample is critical throughout the research process. At the beginning, when a researcher is choosing the size s/he wants their sample to be, determining how many participants to include is an important, yet sometimes tricky process. At the conclusion of the research process when a reviewer is evaluating the validity of claims made based on data presented, the reviewer must evaluate the sample size presented against conclusions drawn. The goal of this paper is to assist researchers and reviewers in consistently determining and evaluating sample size.

ACKNOWLEDGMENTS

We thank Peter Barnett, Emily Matthews, Keanau Ormson, Subina Saini, Kimberly Shappell, Dane Smith, Natalie Smoot, Justin Stephens, and especially Jenna Derrah for assistance with data collection, Cheng Guo, Byron Lowens Vivian Motti, Richard Pak and especially Bart Knijnenburg for their helpful comments on the manuscript. We also thank Julia Sharp and Bart Knijnenburg for assistance with data analysis. Finally, we thank Kathy Baxter, Catherine Courage, and Audrey Giourard for the inspiration for this study. This work was supported by NSF grants 1513875, 1314342, 1117860, and 1228364.

REFERENCES

1. ACM SIGCHI 2015. Guide to a Successful Paper or Note Submission.
2. ACM SIGCHI 2015. Papers Versus Notes Whats the Difference.
3. Herman Aguinis and Erika Harden. 2009. Sample size rules of thumb: evaluating three common practices. In *Statistical and methodological myths and urban legends: doctrine, verity and fable in the organizational and social sciences.* , Charles Lance and Robert Vandenberg Eds., New York, NY, 267-286.
4. William Albert and Thomas Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes.
5. Peter Bacchetti. 2002. Peer review of statistics in medical research: the other problem. *BMJ: British Medical Journal 324*, 7348, 1271.
6. Peter Bacchetti. 2010. Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine 8*, 1, 1-7. http://dx.doi.org/10.1186/1741-7015-8-17.
7. Peter Bacchetti, Steven G. Deeks, and Joseph M. McCune. 2011. Breaking Free of Sample Size Dogma to Perform Innovative Translational Research. *Science Translational Medicine 3*, 87 (2011-06-15 00:00:00), 87ps24-87ps24. http://dx.doi.org/10.1126/scitranslmed.3001628.
8. Peter Bacchetti, Charles E. McCulloch, and Mark R. Segal. 2008. Simple, Defensible Sample Sizes Based on Cost Efficiency. *Biometrics 64*, 2, 577-585. http://dx.doi.org/10.1111/j.1541-0420.2008.01004_1.x.
9. Louise Barkhuus and Jennifer A. Rode. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA2007), ACM, 2180963. http://dx.doi.org/10.1145/1240624.2180963.
10. Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding Your Users: A Practical Guide to User Research Methods*. Morgan Kaufmann.
11. Simone Borsci, Robert D. Macredie, Julie Barnett, Jennifer Martin, Jasna Kuljis, and Terry Young. 2013. Reviewing and Extending the Five-User Assumption: A Grounded Procedure for Interaction Evaluation. *ACM Trans. Comput.-Hum. Interact. 20*, 5, 1-23. http://dx.doi.org/10.1145/2506210.
12. Paul Cairns. 2007. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1* British Computer Society, 195-201.
13. Ed H. Chi. 2011. On the importance of Replication in HCI and Social Computing Research. In *BLOG@CACM*.
14. Jacob Cohen. 1962. The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology 65*, 3, 145-153. http://dx.doi.org/http://dx.doi.org.libproxy.clemson.edu/10.1037/h0045186.
15. Paul D Ellis. 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
16. Barney G Glaser and Anselm L Strauss. 2009. *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers.
17. Henry A Glick. 2011. Sample Size and Power for Cost-Effectiveness Analysis (Part 2). *Pharmacoeconomics 29*, 4, 287-296.
18. Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy2008), ACM, 1357074, 111-120. http://dx.doi.org/10.1145/1357054.1357074.
19. Scott A. Hale. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada2014), ACM, 2557203, 833-842. http://dx.doi.org/10.1145/2556288.2557203.
20. Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences 33*, 2-3, 61-83. http://dx.doi.org/doi:10.1017/S0140525X0999152X.
21. Wonil Hwang and Gavriel Salvendy. 2010. Number of people required for usability evaluation: the 10±2 rule. *Commun. ACM 53*, 5, 130-133. http://dx.doi.org/10.1145/1735223.1735255.
22. Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* ACM, 1105-1114.
23. Helena Chmura Kraemer and Christine Blasey. 2015. *How many subjects?: Statistical power analysis in research*. Sage Publications.
24. Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario,

Canada2014), ACM, 2556969, 3553-3562. http://dx.doi.org/10.1145/2556288.2556969.

25. Jakob Nielsen. 1994. Estimating the number of subjects needed for a thinking aloud test. *International journal of human-computer studies 41*, 3, 385-397.

26. R Nuzzo. 2014. Statistical errors: P values, the 'gold standard'of statistical validity, are not as reliable as many scientists assume. *Nature 506*, 150, 52.

27. Robert J. Ploutz-Snyder, James Fiedler, and Alan H. Feiveson. 2014. Justifying small-n research in scientifically amazing settings: Challenging the notion that only "big-n" studies are worthwhile. *Journal of Applied Physiology*(2014-01-09 22:33:40). http://dx.doi.org/10.1152/japplphysiol.01335.2013.

28. Jenny Preece, Helen Sharp, and Yvonne Rogers. 2015. *Interaction Design-beyond human-computer interaction*. John Wiley & Sons.

29. Daniel Reed and Ed H. Chi. 2012. Online privacy; replicating research results. *Commun. ACM 55*, 10, 8-9. http://dx.doi.org/10.1145/2347736.2347739.

30. Robert Rosenthal. 1965. The volunteer subject. *Human relations 18*, 4, 389.

31. Albrecht Schmidt and Tovi Grossman. 2014. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM, Toronto, Ontario, Canada, 4206.

32. Wendie Wulff and Dick E Mahling. 1990. An assessment of HCI: issues and implications. *ACM SIGCHI Bulletin 22*, 1, 80-87.