Have I Got the Power? Analysing and Reporting Statistical Power in HRI

MADELEINE E. BARTLETT, University of Waterloo, Canada and CRNS, University of Plymouth, United Kingdom
C. E. R. EDMUNDS, Queen Mary, University of London, United Kingdom
TONY BELPAEME, IDLab – imec, Ghent University, Belgium
SERGE THILL, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

This article presents a discussion of the importance of power analyses, providing an overview of when power analyses should be run in the context of the field of Human-Robot Interaction, as well as some examples of how to perform a power analysis. This work was motivated by the observation that the majority of papers published in the proceedings of recent HRI conferences did not report conducting a power analysis; an observation that has concerning implications for many conclusions drawn by these studies. This work is intended to raise awareness and encourage researchers to conduct power analyses when designing research studies using human participants.

CCS Concepts: • **General and reference** → **Surveys and overviews**; **Computing standards, RFCs and guidelines**; **Reference works;**

Additional Key Words and Phrases: Reporting practices, power, methodology, best practice

ACM Reference format:
Madeleine E. Bartlett, C. E. R. Edmunds, Tony Belpaeme, and Serge Thill. 2022. Have I Got the Power? Analysing and Reporting Statistical Power in HRI. *ACM Trans. Hum.-Robot Interact.* 11, 2, Article 16 (February 2022), 16 pages.
https://doi.org/10.1145/3495246

1 INTRODUCTION

In the field of Human-Robot Interaction, we typically run experiments to see how human participants react to robots in some way [Hoffman and Zhao 2020]. For instance, this might involve exploring how certain robot behaviours affect people's perceptions of that robot [Johanson et al. 2019; Winkle et al. 2021], or whether robots can facilitate learning of a second language [Vogt et al. 2019; Wallbridge et al. 2018], or even how children play with them [Boccanfuso et al. 2016]. However, unlike (most) robots, people can be highly unpredictable. In other words, they are noisy. Given an

Authors' addresses: M. E. Bartlett, University of Waterloo, Waterloo, ON, Canada, N2L 3G1, CRNS, University of Plymouth, Plymouth, United Kingdom, PL4 8AA; email: madeleine.bartlett@uwaterloo.ca; C. E. R. Edmunds, Queen Mary, University of London, London, United Kingdom; email: ceredmunds@gmail.com; T. Belpaeme, IDLab–imec, Ghent University, Ghent, Belgium, B-9052; email: Tony.Belpaeme@UGent.be; S. Thill, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, 6525 HR, The Netherlands; email: serge.thill@donders.ru.nl.

2573-9522/2022/02-ART16 $15.00
https://doi.org/10.1145/3495246

identical task on two separate occasions, they are unlikely to complete it in precisely the same way. Thus, to determine whether any observed effects in our experiments are meaningful, rather than just noise, we use statistical methods, most often **Null Hypothesis Statistical Testing (NHST)**.

In NHST, we confirm our hypothesis by statistically comparing our results against a hypothesis of no effect [Pernet 2015]. What this means is that tests of statistical significance are a way of deciding between two possible explanations for an observed effect. One explanation, referred to as the null hypothesis, is a statement that there is no difference or relationship between phenomena or populations, and that the observed effect occurred by chance. However, the alternative hypothesis posits that the effect observed in a sample reflects the effect that exists in the general population. Importantly, NHST rests on the assumption that the null hypothesis is true and determines how likely it is that the observed effect would have occurred if there was no effect in the general population.

To illustrate, imagine you are building a robot for conversational therapy and need to decide on the appearance of the robot. You hypothesise that a robot that has more human-like features will be trusted more than one that is more robot-like. You spend several weeks in the workshop building a human-like and a robot-like robot. You bring in participants and randomly assign them to one of two conditions. One half of your participant group interacts with a human-like robot, the other half of the group interacts with a more robot-like robot. After the interaction you measure trust using a questionnaire. The null hypothesis is that there is no difference in trust between the two conditions: Having a robot that has more human-like features does not increase or decrease trust compared to a robot that appears more robot-like. Using the results and details of the study, NHST methods can be used to calculate how likely it would be that an effect at least at large as the one you found would occur by chance.

When it comes to experimental research using NHST, there are two types of error of which researchers have to be aware. Type I errors are false positives; the null hypothesis is rejected in favour of the alternative hypothesis when the alternative is actually false [Field 2016]. In contrast, Type II errors are false negatives; failing to reject the null hypothesis when the alternative hypothesis is true [Field 2016]. To ensure that our conclusions are valid—and that the study is telling us something meaningful—it is important that researchers control for these two types of error.

Controlling for Type I errors in statistical analysis is fairly straightforward. The probability of a Type I error is known as α (alpha). When conducting statistical significance tests, researchers can control for the probability of a Type I error by setting an acceptable α or significance level [Lieber 1990; Neyman and Pearson 1928]. In practice, researchers control for Type I errors by only accepting statistical results as "significant" if there is a very small chance (e.g., $p < 5\%$) that the result is a false positive. When conducting statistical analyses, tests of significance report a p-value, which denotes the probability of observing results at least as extreme as observed, if the null hypothesis is true. The α level acts as a threshold such that p-values that are smaller than α are considered "significant"; i.e., there is a low probability that the result was a false positive. In most cases, the accepted α level is $\alpha = 0.05$ (5% chance of a Type I error) [Cohen 1988].

Controlling for Type II errors is, arguably, slightly more complicated. The probability of a Type II error (false negative) is commonly denoted as β (beta). This value is used to calculate the **power** of a study such that power $= 1 - \beta$ [McCrum-Gardner 2010]. To ensure that a study has a low risk of producing a false negative finding, studies must be designed to have sufficient power to find a *meaningful* difference. That is, as researchers, we want to avoid conducting studies that are either underpowered or overpowered, as both tend to produce exaggerated and misleading results. As with the alpha level, a researcher can adjust the power value to reflect what they consider an acceptable probability of a Type II error in the statistical tests for their study. Cohen [1988] reasoned that a good balance between α and β would be to have a 5% chance of a Type I error and

a 20% chance of a Type II error (i.e., a power of 80%; power = $1 - \beta$). These values have since been generally accepted as a good default when conducting behavioural research.

Key Terms and Concepts

Null hypothesis
A statement that there is no actual relationship between variables and that any observed effect is due to chance.

Alternative hypothesis
A statement that a difference or effect is not due to chance, suggesting a relationship between variables.

Significance (p-value)
The p-value denotes the probability of observing results at least as extreme as observed if the null hypothesis is true. A smaller p-value indicates that there is stronger evidence in favour of the alternative hypothesis. Significance is indicated by this value being lower than a predefined cut-off (most commonly 0.05 or 5%).

Type I and II errors
Type I and II errors are concerned with either rejecting or accepting the null hypothesis.

A Type I error occurs when we reject the null hypothesis when it is true. It is otherwise referred to as a false positive and is captured by the significance level (p-value) of a test.

A Type II error, however, is when we accept the null hypothesis when it is actually false (i.e., the alternative hypothesis is true). It is therefore referred to as a false negative.

Power
The power of a test is the probability of not making a Type II error. In other words, it measures the ability of a test to correctly reject the null hypothesis.

The most commonly accepted minimum level of power is 80%. If a test has 80% power, then it means that the test has an 80% chance of detecting a difference of a given effect size if such a difference exists. Power is linked to the sample size of a study in that a larger sample size will increase power.

Effect size
Effect size quantifies the difference between groups. It is therefore thought of as indicating the effectiveness of a treatment or experimental condition.

So, how does one design a study to control how probable it is that they will get a false negative (Type II error) result? By performing a power analysis to inform the design of a study, specifically, the sample size. While there are a number of factors that influence the probability of Type II errors (e.g., α, effect size, sample size, and whether the statistical test used is one- or two-tailed [Lieber 1990]), one of the few factors that can be actively controlled by a researcher is sample size [Lieber 1990]. Therefore, power calculations are conducted to establish how many participants a researcher needs to recruit for their study to have a good chance of detecting an effect of the expected size. In the case of small effect sizes, larger sample sizes are generally required to ensure an intended power. From a planning perspective, power analyses in these cases allow researchers to realistically consider whether enough participants can be recruited or whether a different

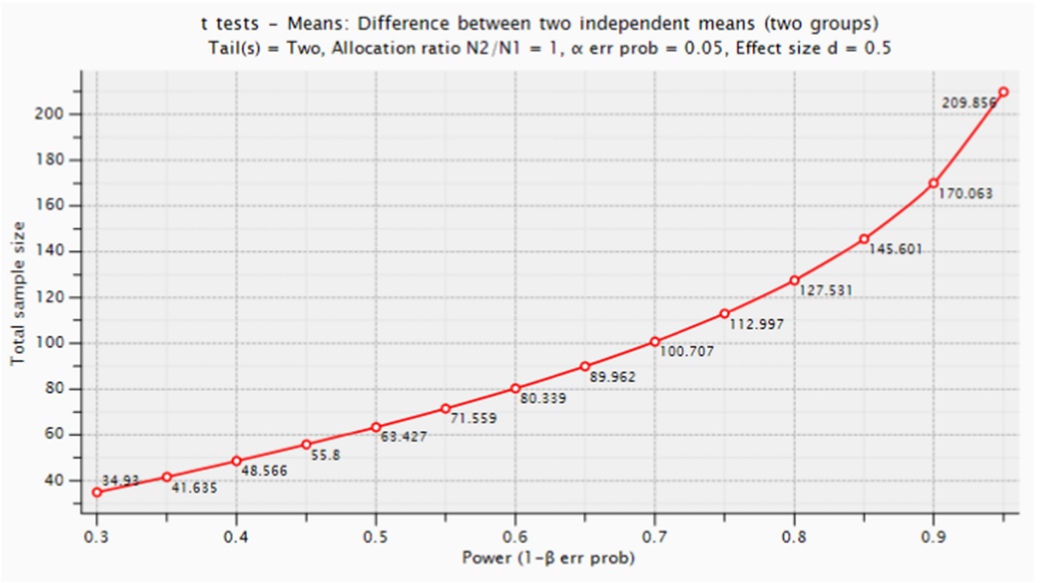

Fig. 1. Plot generated using G*Power, illustrating the positive relationship between sample size and power. This plot was generated for a hypothetical study using a t-test to test the difference between two independent group means, where the chosen alpha level is $\alpha = 0.05$ and the estimated effect size is $d = 0.5$. Labels on the plot are the sample size for each value of power at intervals of 0.05.

recruitment method (e.g., online vs. in-person) or study design (within- vs. between-subjects) should be adopted. However, detecting a large effect size while achieving the same level of power requires fewer participants. Here, power analyses allow researchers to ensure that they do not spend valuable resources recruiting more participants than necessary.

The goal of behavioural research is to conduct studies that tell us something about the general population. For example, in an HRI context one might want to provide evidence that a robot tutor is *significantly* more effective at promoting learning than a human tutor. To test this, a study could be run comparing the learning gains of two groups: one that is taught by the robot tutor and one taught by the human tutor. A t-test comparing group means produces a p-value of $p = 0.03$. Because we chose an alpha level of $\alpha = 0.05$, we consider this result to be significant, and it effectively demonstrates that, if the null hypothesis were true, then a result this extreme would occur only 3% of the time [Ferreira and Patino 2015; Jhangiani et al. 2015].

Now let us say that, to achieve a power level of power $= 0.8$ and detect an effect size of $d = 0.5$ the study needed 64 participants in each group (128 participants total, as indicated by Figure 1). Unfortunately, only 30 participants were recruited in each group (60 in total). By referring back to Figure 1, we can see that this resulted in the study only achieving a power level of roughly 0.475, i.e., the study was underpowered. This changes how the findings are interpreted. Studies that are underpowered are more susceptible to random variation in sample means. One consequence of this is that there is a higher chance of producing a large effect size and of reaching statistical significance [Gelman and Weakliem 2009]. These large, significant effects, however, come with a high degree of uncertainty due to the low statistical power, and it is important that this uncertainty be taken into account. That is, rather than concluding that having a robot tutor produces significantly greater learning gains than having a human tutor, it would be more

Table 1. Number of Papers in Main Proceedings of HRI 2020 and RO-MAN 2019, Number of Papers Reporting Experiments, and Number of Experiment Papers Reporting Power Analyses

Conference	N Papers	N Experiments	N Reporting Power
HRI 2020	66	34	5
RO-MAN 2019	187	34	0

Table 2. Table of the Reported Suggested Sample Sizes and Actual Sample Sizes for Papers Reporting *a priori* Power Analyses

	Required Sample Size	Actual Sample Size
Paper 1	<250	49
Paper 2, study 1	32	47
Paper 2, study 2	32	51
Paper 2, study 3	32	72
Paper 3	73	99

accurate to state that "our results suggest that having a robot tutor may result in better learning than having a human tutor for this task, but the 'true' effect may be smaller than that found, and additional studies with greater power are needed to establish a better estimate of the true effect."

Power analyses allow us to calculate the number of participants needed while controlling for both α and β for us to be confident in our results. They are also extremely valuable to readers so the results being presented in a research paper can be properly interpreted. However, there seems to be a lack of power analyses reported in papers in the field of HRI. To examine whether or not this is the case, the next section provides a brief look at how often papers published in the field of HRI include reports of power calculations.

2 THE STATE OF AFFAIRS

To illustrate that there is a need within the field of HRI to encourage the use and reporting of power analyses, we investigated the number of papers reporting power analyses in two recent conferences. Specifically, due to the large amount of noise in studies using human participants and thus, the importance of statistical methods, we exclusively focused on papers that reported running an experiment using human participants, where the goal was to examine the effect of some independent variable on a human-factors dependent variable (e.g., perceptions of robots, task performance metrics). To be considered an experiment for the purposes of this analysis, the study needed to be comparing the effect of at least two conditions. Pilot studies were not included. We first collected all the papers published in the main conference proceedings of the *2020 ACM/IEEE International Conference on Human-Robot Interaction* (HRI) and the *2019 IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN) (see Table 1). In total, there were 68 papers published across both conferences that reported experimental studies using human participants, 5 of which reported conducting a power analysis. Of these, 3 papers reported conducting *a priori* power analyses to calculate the required sample size. For one of these papers, three studies were conducted and power analyses run for each study. The reported required sample sizes and actual sample sizes for all 5 studies (across three papers) are shown in Table 2. The remaining 2 papers reported conducting *post hoc* power analyses.

Given that all 68 of these studies should have reported power, this finding indicates that power analyses are under-reported in HRI. In an effort to address this, this article is intended as a resource

for highlighting the importance and utility of power analyses, and providing examples of when and how one might conduct power analyses, for a number of different study designs found in HRI.

3 WHEN TO REPORT POWER

As a general rule, if an experiment involves human participants and the use of NHST, then a power analysis should be used in the planning stages to establish the sample size needed to achieve the researcher's chosen power level.

The wide range of study designs involved in HRI means that this "rule of thumb" approach could be too restrictive. In such instances, the following checklist might be a useful resource. If a study involves:

(1) recruiting human participants, AND
(2) comparing the effects of at least two conditions, OR comparing a sample mean to a population mean, AND
(3) a dependent variable that is a measure of human behaviour or cognition, AND
(4) tests of statistical significance,

then a power analysis should be conducted to establish the required sample size and reported to inform the interpretation of the results.

There are studies that do not require power analyses. These include; studies where NHST is substituted with Bayesian statistics [Correll et al. 2020], descriptive studies [Aggarwal and Ranganathan 2019], and exploratory research [Kraemer and Blasey 2015].

4 UNDER- AND OVERPOWERED STUDIES

4.1 Underpowered Studies

It is worth noting that we are not arguing that every study must ensure adequate power but simply that the power analysis should be reported. There may be good reasons to go ahead with an underpowered study; for example, if it focuses on a population from which obtaining a sufficient sample size is not realistic (such as persons of advanced age). Underpowered studies can also be useful in raising awareness around a new research avenue, or as a stepping-stone for larger studies.

At the same time, running an underpowered study might also raise ethical concerns, because it requires investments from participants (even if it is just time) into a study whose ability to generate insights might be limited. In particular, this concern exists when dealing with vulnerable participant groups, children, and so on; in other words, a situation in which underpowered studies are likely because, as mentioned above, it may simply not be possible to recruit sufficient numbers from this population. Since this is an ethical concern, it falls within the remit of the relevant ethical committee or authority to assess it against the possible benefits of the study. A power analysis can help the committee reach the appropriate decision.

If there are no concerns preventing the study from being run, then it can also be published. In that case, it remains important for authors to provide the power analysis, acknowledging the underpowered nature of the study (including an explanation of why this was unavoidable) and ensure that conclusions are phrased accordingly; in particular, avoiding overly assertive or strong claims. At the same time, it is also important for editors and reviewers to realise that there is no "magical" power threshold with an automatic rejection on one side of it. As with p-values, confidence intervals, and so on, such measures exist on a continuum, and it is ultimately up to the reader to decide what confidence they have in the results. However, for readers to be able to make an informed decision, they need to be provided with the necessary descriptors, including a power analysis.

To illustrate how such instances could be addressed in publications, we suggest that, if authors find that the power analysis done during the planning fails to suggest an unrealistic or unattainable sample size, and a new, smaller, sample size is selected, then researchers can use a plot similar to that presented in Figure 1 to calculate the power level achieved with this sample. This power level can then be reported alongside a brief explanation of why the "preferred" sample size could not be recruited.

4.2 Overpowered Studies

Recently, especially during the COVID-19 pandemic, online data collection, also known as "crowd sourcing,"[1] has proven very popular to collect experimental data. Many studies, not only in HRI [Berinsky et al. 2012; Buhrmester et al. 2016], now rely on online crowd sourcing, as data can be collected with lower effort, often at lower cost, in a shorter time and from a much wider population. This is a promising development, especially when we are prevented from letting participants interact with real robots. However, as recruiting more participants increases the power of a study, it brings with it the problem of potentially *overpowering* a study [Hochster 2008].

A large number of participants, with hundreds or even thousands of participants not being an exception in online HRI studies, increases the statistical power of a study. However, it also is more likely that low p-values will be observed [Sellke et al. 2001]. In other words, increasing the number of participants increases the probability of getting a significant result even for very small differences between conditions. So, while your results are statistically significant, they might well be scientifically uninteresting.

Overpowered studies are not wrong in a statistical sense, but they are ethically questionable when only significance is reported without also reporting the effect sizes and alongside a discussion on the relevance of the effect found. In addition, overpowered studies waste resources. Time, money, and participants are valuable to most of us, and correctly assessing the number of participants needed to answer a research questions means that no more resources are used than strictly needed.

5 HOW TO CALCULATE POWER

Numerous tools exist for calculating sample size, including G*Power [Faul et al. 2007], PASS [Kaysville UT: NCSS. 2018], SAS [SAS Institute 2004], the pwr package for R [Champely et al. 2020], and the Power and Sample Size website [HyLown Consulting LLC HyLown Consulting LLC]. The following examples were conducted using G*Power:

To run a power analysis, we first need to collect/define a few key pieces of information:

(1) What statistical test will be used?
(2) What is the chosen value for α?
(3) What is the size of the effect that we predict we will find?

Question 1: For question 1, we consider only the statistical test planned for testing the main research question. So, if a study is looking at the effect of robot tutor vs. human tutor on learning, then the main test might be comparing the groups' average test scores. If run as a between-subjects study, then the test would be an Independent Samples T-test or 1-way ANOVA. Alternatively, such a study could be run using a within-subjects design (thus reducing the required sample size), which would require a Repeated-Measures ANOVA [Field 2016].

[1]With Amazon Mechanical Turk or Prolific being two popular commercial platforms for setting up paid online studies.

Question 2: Answering question 2 is, in most cases, a simple task of deciding whether or not to stick with the standard $\alpha = 0.05$ significance level (although using a lower α has been argued to improve replication [Gibson 2021]).

Question 3: There are a couple of different ways to answer question 3. Effect size quantifies the magnitude of an experimental effect. In an experiment comparing an experimental group to a control group, the effect size quantifies the difference between the two groups or means [Coe 2002; McMillan et al. 2002]. In correlation analyses with two or more variables, or studies where there are more than two groups, the effect size measures the strength of the association between variables [McMillan et al. 2002]. They can be thought of as measuring the correlation between the effect and the dependent variable. Estimating the effect size is arguably the most difficult step in this process. One way of obtaining an estimate of the effect size is by looking at existing, similar research and calculating an average effect size that represents an estimate of the population effect size. However, this approach is subject to potential bias introduced by the "file drawer effect" [Anderson et al. 2017], whereby the preference for publishing statistically significant results has likely resulted in the published effect sizes being an over-estimation of the actual population effect size. A direct consequence of this is that future studies that use this method for estimating effect sizes may be underpowered. Some have therefore advised that researchers take a conservative approach when using this method [Correll et al. 2020].

An alternative approach is to conduct a pilot study before conducting the full study. This approach, however, does come with its own problems, one being that the variability of effect sizes found during small pilot studies will be large [Brysbaert 2019]. Because of this, researchers may be less inclined to conduct a full study if the pilot study reveals a small or insignificant effect, but pilot studies that reveal large effects might be over-estimations and also lead to underpowered full studies [Albers and Lakens 2018; Anderson et al. 2017].

A third option is to use suggested effect size values based on what test we plan to use and whether we expect to find a small, medium, or large effect [Cohen 1992; Ferguson 2016]. For example, if our test is a comparison of independent means (i.e., independent t-test) and we want to be able to detect a medium effect size, then Cohen [1992] suggests an estimated effect size of $d = 0.5$. It should be noted, at this point, that Cohen's definitions of what constitutes a "small," "medium," and "large" effect size are inconsistent across different measures of effect size. A recent paper by Correll et al. [2020] provides an in-depth discussion of this. We therefore recommend caution when taking this approach. In our examples below, we demonstrate one of the recommendations made in Correll et al. [2020]. That is, to convert between a standard effect size, e.g., η^2, and other measures, rather than relying on the different definitions provided by Cohen [1992].

To our knowledge, there is currently no "perfect" approach to estimating effect size for power analyses. The best recommendation we can provide, therefore, is to use caution and utilise conservative effect size estimates. The following sections provide examples of how to calculate required sample size for a few different studies.

5.1 Example 1 - Independent Samples T-test

For these demonstrations of power calculations, we use the example study of comparing the effect of a robot vs. a human tutor on learning gains, and we use G*Power to perform the power analysis. In this first example, the hypothetical study is simply intended to compare the effect of tutor on learning. Participants are taught either by a robot or a human tutor and then tested at the end of the learning phase. The average test scores of each group will be compared using an independent t-test.

The effect size for the power analysis is obtained by calculating an estimate of the population effect size based on previous, similar studies. This reveals an average effect size of $d = 0.46$. The

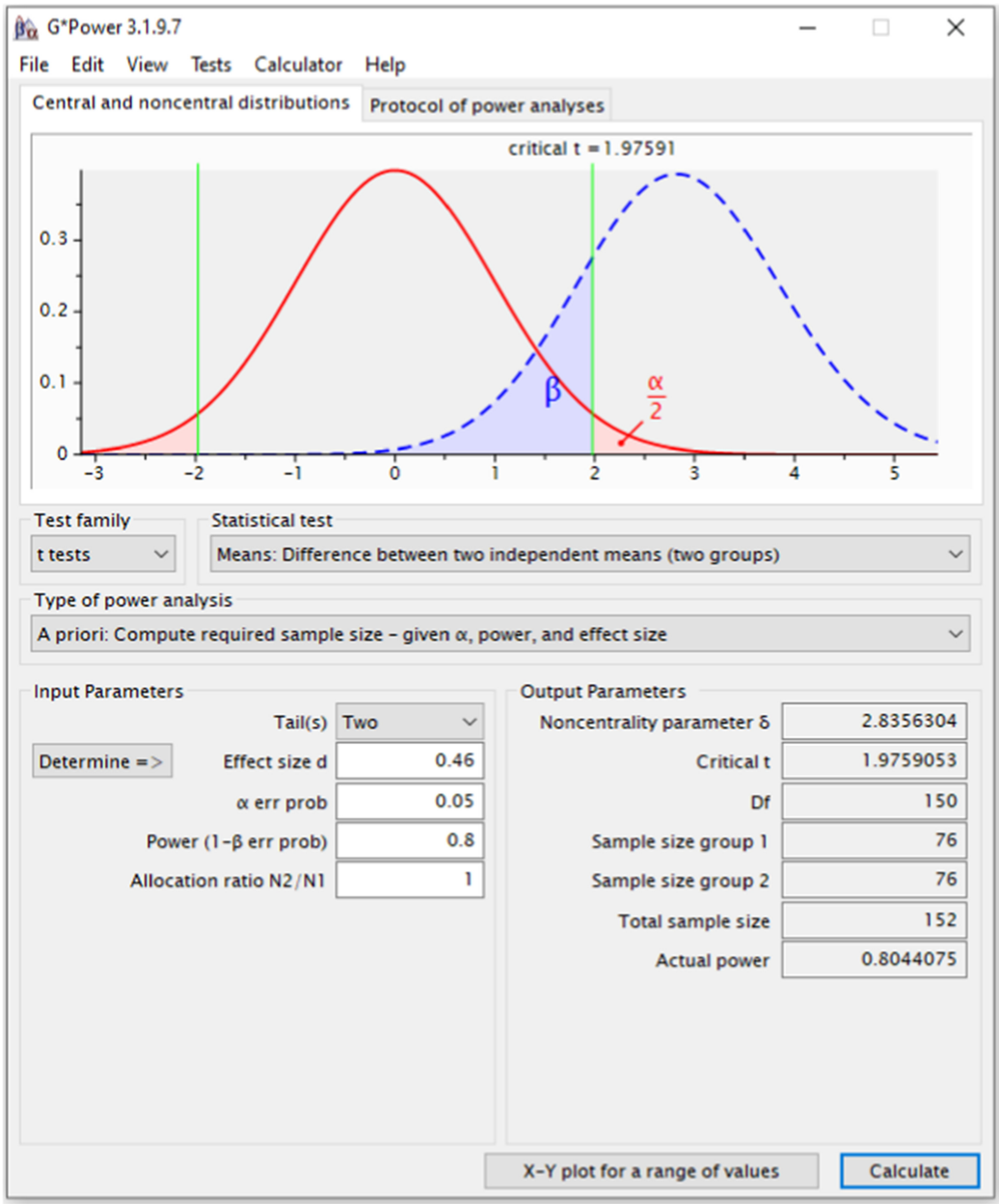

Fig. 2. Screenshot of G*Power. Calculating required sample size for test comparing two independent groups.

chosen alpha level is $\alpha = 0.05$, and power level is $1 - \beta = 0.8$. The G*Power software calculates that the required sample size is 152 (76 in each group) (see Figure 2).

This is a rather large required sample size and might not be achievable. One way to reduce this requirement is to instead use a within-subjects design that reduces the total required sample size to 40 (see Figure 3).

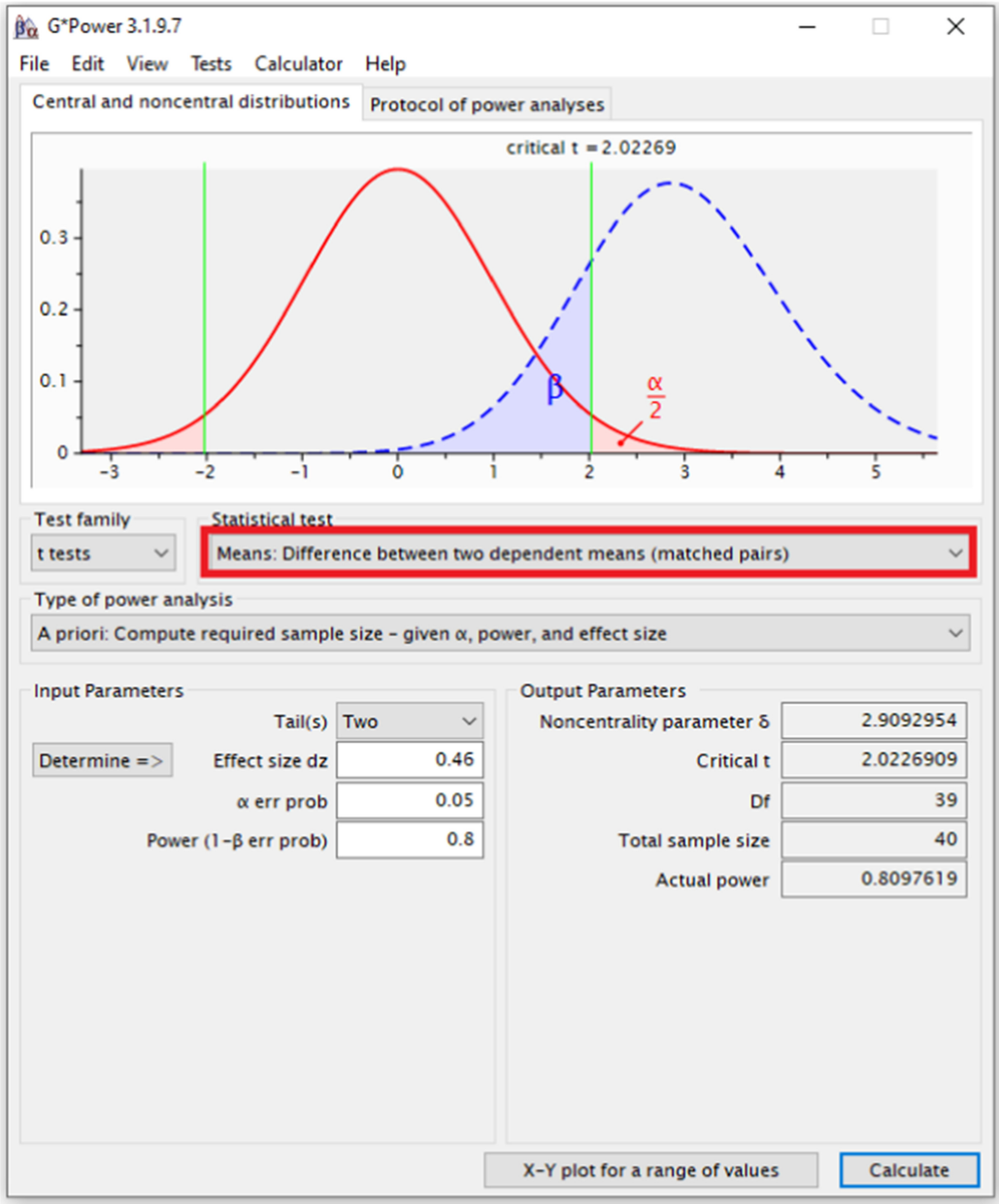

Fig. 3. Screenshot of G*Power. Calculating required sample size for test comparing two dependent groups. These can be matched pairs or within-subjects. Highlighted is the Statistical test box, where users can change whether the test is for a between- or within-subjects design.

5.2 Example 2 - 2-way ANOVA

To demonstrate a power analysis for a 2-way ANOVA the study needs to involve two independent variables. Let us therefore imagine that we want to look at the effect of both tutor (robot vs. human) and subject difficulty (easy vs. hard) on learning gains. The study also uses a

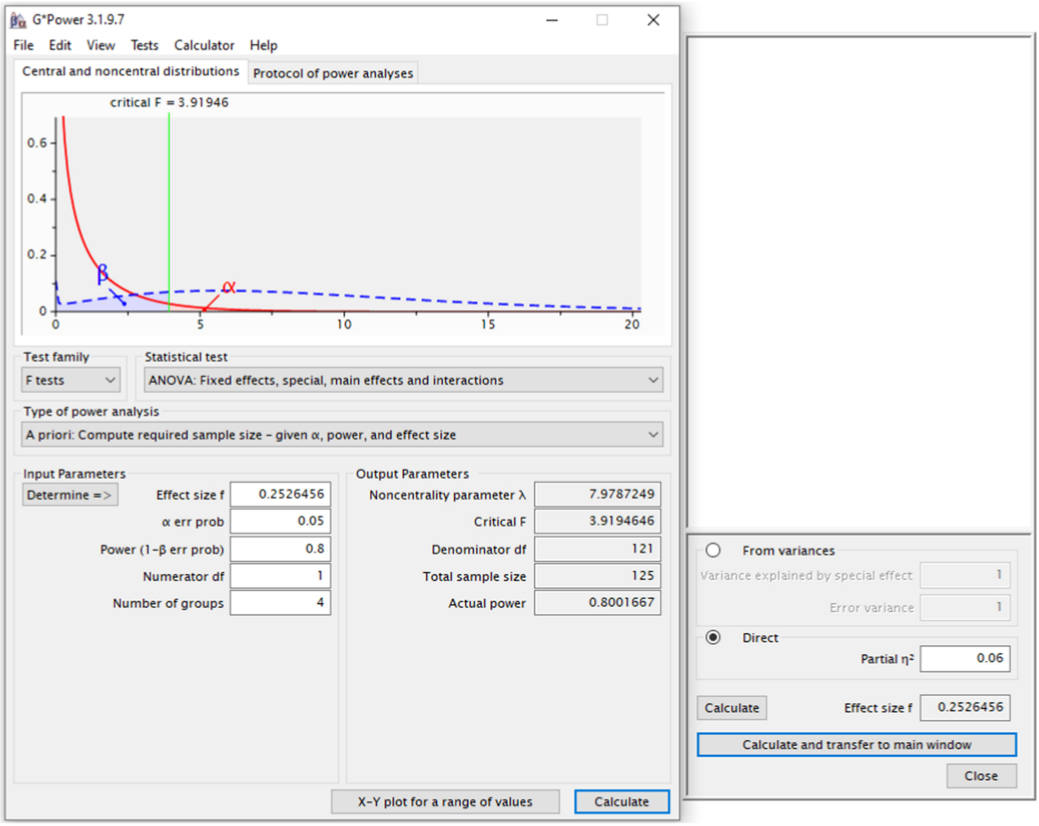

Fig. 4. Screenshot of G*Power. Calculating required sample size for test comparing four independent groups.

between-subjects design, requiring four independent groups. Here, following one of the recommendations from Correll et al. [2020], we calculate a "medium" effect size based on Cohen's medium value for $\eta^2 = 0.06$ [Cohen 1988]. G*Power can do this and gives the equivalent $f = 0.253$. As in the previous example, we calculate the required sample size using an alpha level of $\alpha = 0.05$, and power level is $1 - \beta = 0.8$. G*Power gives the required sample size of 125 participants (31 or 32 per group) (see Figure 4).

5.3 Example 3 - Repeated Measures ANOVA

As a third example, let us propose a study using a within-subjects design. The study is the same as the previous example, where we examine the effect of tutor (robot vs. human) and task difficulty (easy vs. hard) on learning gains, but participants see all four conditions. The estimated effect size is again $f = 0.253$, and we chose the parameters $\alpha = 0.05$ and $1 - \beta = 0.8$. There are no between-subjects factors, so the study has one group and four measurements. A pilot study can be used to get an estimate of the correlation between repeated measures. In this case, we will imagine this shows a correlation of roughly $\rho = 0.5$ and that the sphericity assumption was not violated (therefore $\epsilon = 1$). This analysis reveals that the sample size required to ensure the chosen power level is 23 (see Figure 5).

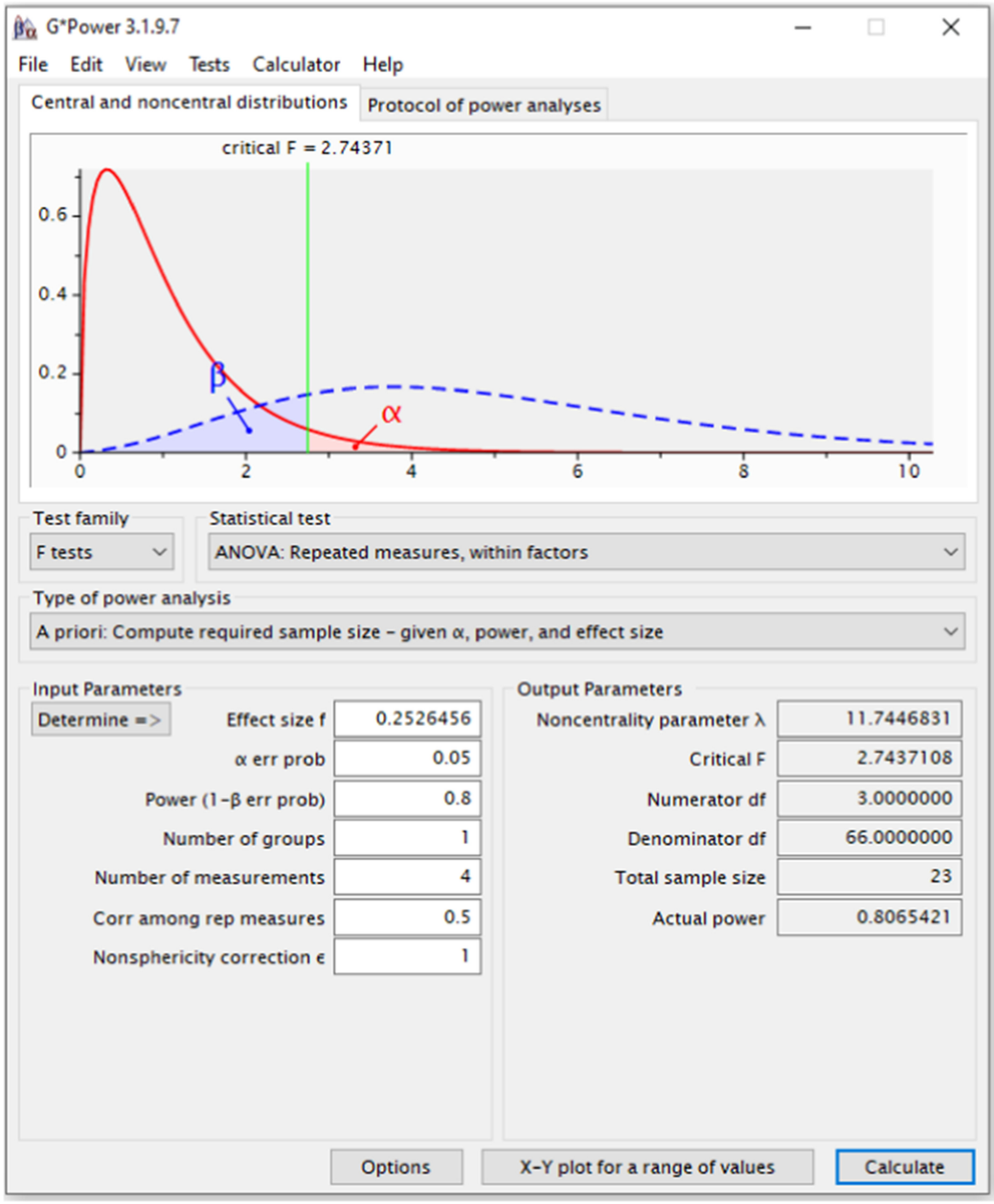

Fig. 5. Screenshot of G*Power. Calculating required sample size for a 2 × 2 within-subjects repeated measures test.

6 A NOTE ON POST HOC POWER

The examples we have provided here are examples of *a priori* power analyses—calculations done before data collection to inform study design. Another type of power analysis that can be done is post hoc. Post hoc power analyses are conducted after the study has been conducted. There

are a few ways to approach conducting post hoc power analyses. The first can be thought of as simply doing the power analyses after the study. In this instance, the power analysis is conducted using the same method as an *a priori* analysis. That is, using the chosen alpha and target power values, an estimated effect size, and calculating how many participants would have been required to achieve the target power. This can then be compared with the actual sample size to assess the study. Alternatively, one can use the chosen alpha level, estimated effect size, and actual sample size to estimate the achieved power level. Note that in both of these cases it is the *estimated effect size* and not the observed effect size that results from the study, which is used in the analysis. These methods can be useful for interpreting the study's results, similar to *a priori* power calculations and for providing a possible explanation for inconclusive results. For instance, if a study did not reveal any effect, then this kind of post hoc analysis might suggest that the study was underpowered and that a second, better powered study should be conducted.

Another approach that is often seen is to use the actual sample size and the observed effect size to calculate power. This has also been referred to as "observed power" or "retrospective power" [O'Keefe 2007]. However, it has been argued that this approach is not as meaningful as once thought. The purpose of a power analysis is to determine the likelihood that a chosen statistical test will detect an effect at least as large as the "real" or population effect, assuming there is one. However, power analyses that use the observed effect size are determining the likelihood that the test will produce a statistically significant result, assuming that the population effect is the same as the observed effect [Levine and Ensom 2001; O'Keefe 2007; Sebyhed and Gunnarsson 2020; Thomas 1997]. In the first instance, this implies that a very strong assumption is being made. Additionally, it has been proven that there is a 1-1 relationship between p-values and this kind of observed power statistic such that tests with larger p-values always have low "observed" power [Hoenig and Heisey 2001]. Post hoc power analyses, therefore, offer little additional insight in the case of non-significant results.

7 CONCLUSION

This article presents a discussion of the importance of power calculations in response to the observation that power seems to be under-reported in the field of HRI. The first step to addressing this is to encourage researchers to perform power calculations and report these calculations in their papers.

One possible reason that power analyses are under-reported might be that the majority of researchers have not received sufficient training on the use and importance of power analyses and sample size calculations. This is partly due to the multi-disciplinary nature of HRI. Just as one would not expect a psychologist to have received formal training on how to develop or program a robot, one would not expect someone trained in robotics or software engineering to have also received training on how to design experiments involving human participants. Fortunately, there is a wide range of educational resources available on power analyses and the importance of calculating sample size, including informational and instructional **books** [Field 2016; Hedberg 2017; Jost et al. 2020; Kraemer and Blasey 2015], **articles** [Baxter et al. 2016; Faul et al. 2007; Jones et al. 2003; Lieber 1990; McCrum-Gardner 2010; Prajapati et al. 2010], and **videos** [Center for Open Science 2015; GraphPad Software 2020; Khan Academy 2018; StatQuest with Josh Starmer 2020]. Many of these resources are freely available, and links to the YouTube videos and channels can be found in the reference section of this article.

Another way to educate researchers on power analysis, which may be more beneficial and potentially reach more people, would be to introduce training and educational workshops at more conferences where the focus is on learning skills and techniques rather than on presenting recent research. In HRI, these types of workshops could be run not only to teach researchers about power

analysis, but other skills such as statistical analysis methods, coding for social robotics, and even introductions to newly developed datasets so researchers can have hands-on experience of what a dataset contains and how it might be used.

In terms of the current state of the field, the lack of power analyses in existing research is concerning in that it suggests that the conclusions that have so far been drawn may not be accurate. Another, hugely important, next step then is to replicate. In general, more emphasis on the importance of replications is desperately needed in most, if not all, scientific fields. It is a problem that is difficult to address from within the research community. There are several reasons for this. First, most journals prioritise, or even require, novelty in the studies that they publish. A recent review of over 1,000 psychology journals found that only 3% stated that they welcomed replication studies for publication [Martin and Clarke 2017]. Additionally, it appears that funding is not widely available for replication studies, considering that the Netherlands Organisation for Scientific Research made headlines in 2016 with the first grant programme dedicated to replication studies [Baker 2016]. Thus, encouraging replications is not a trivial issue and requires the coordinated action of researchers, funding agencies, and journals. In the short term however, providing educational resources and ensuring the use of power analyses will enable us to be more confident in the conclusions that we draw from future research.

8 A FINAL NOTE FROM THE AUTHORS

While this article was written to help address the issue of under-reporting of power analyses and provide some educational resources, it is not enough. There are already numerous publications on this topic that are freely available. So, while it is our hope that this article will encourage and empower individual researchers to use power analyses in future studies, this article is also addressed to publication venues and conference organizers as a call to encourage the use and reporting of power analyses in experimental research papers involving human participants.

REFERENCES

R. Aggarwal and P. Ranganathan. 2019. Study designs: Part 2–descriptive studies. *Perspect. Clin. Res.* 10, 1 (2019), 34.

C. Albers and D. Lakens. 2018. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *J. Exper. Soc. Psychol.* 74 (2018), 187–195.

S. F. Anderson, K. Kelley, and S. E. Maxwell. 2017. Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychol. Sci.* 28, 11 (2017), 1547–1562.

M. Baker. 2016. Dutch agency launches first grants programme dedicated to replication. Retrieved from https://www.nature.com/news/dutch-agency-launches-first-grants-programme-dedicated-to-replication-1.20287.

P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 391–398.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* 20, 3 (2012), 351–368.

L. Boccanfuso, E. Barney, C. Foster, Y. A. Ahn, K. Chawarska, B. Scassellati, and F. Shic. 2016. Emotional robot to examine different play patterns and affective responses of children with and without ASD. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 19–26.

Marc Brysbaert. 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* 2, 1, Article 16 (2019). https://doi.org/10.5334/joc.72

M. Buhrmester, T. Kwang, and S. D. Gosling. 2016. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? In *Methodological Issues and Strategies in Clinical Research*, A. E. Kazdin (Ed.). American Psychological Association, 133–139. https://doi.org/10.1037/14805-009

Center for Open Science. 2015. What is statistical power. Retrieved from https://www.youtube.com/watch?v=-ZU7fbvSJ60&ab_channel=CenterforOpenScience.

S. Champely, C. Ekstrom, P. Dalgaard, J. Gill, S. Weibelzahl, A. Anandkumar, C. Ford, R. Volcic, and H. De Rosario. 2020. Basic Functions for Power Analysis. Retrieved from https://github.com/heliosdrm/pwr.

R. Coe. 2002. It's the effect size, stupid. In *British Educational Research Association Annual Conference*. Vol. 12. 14.

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Academic, New York, NY.
J. Cohen. 1992. A power primer. *Psychol. Bull.* 112, 1 (1992), 155.
J. Correll, C. Mellinger, G. H. McClelland, and C. M. Judd. 2020. Avoid Cohen's "small," "medium," and "large" for power analysis. *Trends Cogn. Sci.* 24, 3 (2020), 200–207.
F. Faul, E. Erdfelder, A. Lang, and A. Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Meth.* 39, 2 (2007), 175–191.
C. J. Ferguson. 2016. An effect size primer: A guide for clinicians and researchers. In *Methodological Issues and Strategies in Clinical Research.* American Psychological Association, 301–310. https://doi.org/10.1037/14805-020
J. C. Ferreira and C. M. Patino. 2015. What does the p value really mean? *J. Brasil. Pneumol.* 41, 5 (2015), 485.
A. Field. 2016. *An Adventure in Statistics: The Reality Enigma.* Sage.
A. Gelman and D. Weakliem. 2009. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *Amer. Sci.* 97, 4 (2009), 310–316.
E. W. Gibson. 2021. The role of p-values in judging the strength of evidence and realistic replication expectations. *Statist. Biopharm. Res.* 13, 1 (2021), 6–18.
GraphPad Software. 2020. Calculating Sample Size with Power Analysis. Retrieved from https://www.youtube.com/watch?v=KIRwsYTR62A&ab_channel=GraphPadSoftware.
E. C. Hedberg. 2017. *Introduction to Power Analysis: Two-group Studies.* Vol. 176. Sage Publications.
Howard S. Hochster. 2008. The power of "p": On overpowered clinical trials and "positive" results. *Gastroint. Cancer Res.* 2, 2 (2008), 108.
J. M. Hoenig and D. M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Amer. Statist.* 55, 1 (2001), 19–24.
Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human-robot interaction. *ACM Trans. Hum.-robot Interact.* 10, 1 (2020), 1–31.
HyLown Consulting LLC. Power and Sample size. 2013–2021 Retrieved from http://powerandsamplesize.com/.
R. S. Jhangiani, I. A. Chiang, and P. C. Price. 2015. *Research Methods in Psychology-2nd Canadian Edition.* BC Campus.
D. L. Johanson, H. S. Ahn, B. A. MacDonald, B. K. Ahn, J. Lim, E. Hwang, C. J. Sutherland, and E. Broadbent. 2019. The effect of robot attentional behaviors on user perceptions and behaviors in a simulated health care interaction: Randomized controlled trial. *J. Med. Internet Res.* 21, 10 (2019), e13667.
S. Jones, S. Carley, and M. Harrison. 2003. An introduction to power and sample size estimation. *Emerg. Medic. J.* 20, 5 (2003), 453.
C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig. 2020. *Human-Robot Interaction.* Springer.
Kaysville UT: NCSS. 2018. Power Analysis and Sample Size Software. Retrieved from https://www.ncss.com/software/pass/.
Khan Academy. 2018. Introduction to power in significance tests | AP Statistics | Khan Academy. Retrieved from https://www.youtube.com/watch?v=6_Cuz0QqRWc&ab_channel=KhanAcademy.
H. C. Kraemer and C. Blasey. 2015. *How Many subjects?: Statistical Power Analysis in Research.* Sage Publications.
M. Levine and M. H. H. Ensom. 2001. Post hoc power analysis: An idea whose time has passed? *Pharmacother. J. Hum. Pharmacol. Drug Ther.* 21, 4 (2001), 405–409.
R. L. Lieber. 1990. Statistical significance and statistical power in hypothesis testing. *J. Ortho. Res.* 8, 2 (1990), 304–309.
G. N. Martin and R. M. Clarke. 2017. Are psychology journals anti-replication? A snapshot of editorial practices. *Front. Psychol.* 8 (2017), 523.
E. McCrum-Gardner. 2010. Sample size and power calculations made simple. *Int. J. Ther. Rehab.* 17, 1 (2010), 10–14.
J. H. McMillan, S. Lawson, K. Lewis, and A. Snyder. 2002. Reporting effect size: The road less traveled. In *Annual Meeting of the American Educational Research Association, New Orleans, LA.*
J. Neyman and E. S. Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* (1928), 175–240.
D. J. O'Keefe. 2007. Brief report: Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Commun. Meth. Meas.* 1, 4 (2007), 291–299.
C. Pernet. 2015. Null hypothesis significance testing: A short tutorial. *F1000Research* 4 (2015).
B. Prajapati, M. Dunne, and R. Armstrong. 2010. Sample size estimation and statistical power analyses. *Optom. Today* 16, 7 (2010), 10–18.
SAS Institute. 2004. Getting Started with the SAS Power and Sample Size Application. (2004).
H. Sebyhed and E. Gunnarsson. 2020. *The Impotency of Post Hoc Power.* Dissertation. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-433274.
Thomas Sellke, M. J. Bayarri, and James O. Berger. 2001. Calibration of p values for testing precise null hypotheses. *Amer. Statist.* 55, 1 (2001), 62–71.
StatQuest with Josh Starmer. 2020. Power Analysis, Clearly Explained!!! (2020). Retrieved from https://www.youtube.com/watch?v=VX_M3tIyiYk&ab_channel=StatQuestwithJoshStarmer.

L. Thomas. 1997. Retrospective power analysis. *Conserv. Biol.* 11, 1 (1997), 276–280.

Paul Vogt, Rianne van den Berghe, Mirjam de Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, Ora Oudgenoeg-Paz, Daniel Hernández García, et al. 2019. Second language tutoring using social robots: A large-scale study. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 497–505.

C. D. Wallbridge, R. van den Berghe, D. Hernández Garcia, J. Kanero, S. Lemaignan, C. E. R. Edmunds, and T. Belpaeme. 2018. Using a robot peer to encourage the production of spatial concepts in a second language. In *6th International Conference on Human-Agent Interaction*. 54–60.

K. Winkle, G. I. Melsión, D. McMillan, and I. Leite. 2021. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In *ACM/IEEE International Conference on Human-Robot Interaction*. 29–37.

Received May 2020; revised September 2021; accepted October 2021