Submission

Title Running an HCI Experiment in Multiple Parallel Universes

Authors Pierre Dragicevic, INRIA, Saclay, France
Fanny Chevalier, INRIA, Lille, France
Stephane Huot, Université Paris-Sud, Orsay, France

Keywords

Abstract We experimentally evaluated a haptic touch slider in 8 parallel universes. The results were overall similar but exhibited surprisingly high variability in terms of statistical significance patterns. We discuss the general implications of these findings for empirical HCI research.

Files Submission (PDF, 3.0 MB)

Reviews and Comments

A review indicates for quality, appropriateness, and likelihood to promote discussion. Replies and comments do not include scores, but can add to the discussion. Of course, you can't review your own paper, but you can add a comment or reply to an existing review.

Please note: CHI 2014 alt.chi is an open reviewing forum. The authors and other reviewers will see your name when you add a review, comment, or reply.

Add a review (as an author, you cannot do this)

Add a comment

Steve Szigeti - CIV-DDD, OCAD University, Toronto, Ontario, Canada

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	3 (Knowledgeable)
Conflict of Interest	No, I do not have a conflict
The Review	A clever idea for a paper that will result in excellent discussion. Papers such as this one are important, since they instill a sense of playfulness (whimsy?) to the conference, but nonetheless make some important claims. In this case, the authors' discussion of the role of chance - specifically issues which emerge when evaluating with a relatively small number of participants - is valid and an important consideration in reporting research outcomes. The main negative with the paper stems from its limitations. The authors conduct evaluations in only eight parallel universes, where arguably a proper multiverse experiment would include a minimum of 15, of which three of the selected universes would ideally be quilted, cyclic and brane (as postulated in the work of Brian Greene). However, I strongly recommend acceptance of this work.
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Garth Shoemaker - Google, Mountain View, California, United States

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	3 (Knowledgeable)
Conflict of Interest	No, I do not have a conflict
The Review	I initially took this paper to be merely clever. After reading further I realized that it is not that at all, but is rather insightful and an important contribution to the community. The one fault I found is that the analysis didn't include the experiment from the ninth parallel universe, in which I was a co-author.

Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Peter Pirolli - Palo Alto Research Center (PARC), Palo Alto, California, United States

This is great! There will be a CHI panel on Interaction Science with replication as one foci and I encourage people to come to that.

The has been an alarming number of articles in psychology and elsewhere about these issues (replication; reliability the obsession with p-values). For instance, even the Economist has reported about this: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble?frsc=dg%7Cc

I suspect that experiments reported at CHI are even poorer than ones in these other disciplines because they typically have such low N (are underpowered). For instance, what is the power of an N = 12 experiment (such as the one reported in this paper). You could not reliably detect (with 80% power) that
* men weigh more than women (need N = 46)
* people who like eggs report eating egg salad more often (need N = 28))
* people who like spicy food report eating Indian food more often (need N=26)
[Thanks to Uri Simonshon for generating these results]

Given that many CHI paper have Ns < 50 and report significant effects, is it really the case that all the effects they report are stronger than the genetic difference in body weights between men and women?

reply

Theophanis Tsandilas - Inria, Orsay, France

I have a conflict with the authors, but I would like to add a comment to the discussion opened by this clever work. As the authors replied, the point of the paper is not about sample sizes or power per se.

Consider that a typical within-subjects experiment with 12 participants (like the one reported in the paper that compares two techniques) will often have more power than a between-subjects experiment (as one involving measurements of the weight of women and men) with 40 participants. A simple power analysis with a tool such as G*Power is enough for verifying that. Luckily, HCI experimental methodology is not completely flawed. Also, what seems to be logical or obvious based on our past experience and our interaction with the world is not necessarily easier to show statistically, i.e., with a lower number of samples.

I don't have any concrete data, but (purely based on my own experience and intuition) I believe that the power of typical well-designed HCI experiments (with ~12 participants) that introduce interaction techniques with meaningful, worth-studying and practically (NOT just statistically) significant performance benefits normally have enough power (>80%) due to their interest in high effect sizes. It is true that much of the HCI research focuses more on innovation, i.e., designing new techniques with clear benefits (and consistency among a well-identified group of users) than observing real but weak phenomena.

reply

Thomas Methven - School of Mathematical & Computer Sciences, Heriot-Watt University, Edinburgh, Midlothian, United Kingdom

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	3 (Knowledgeable)
Conflict of Interest	No, I do not have a conflict
The Review	`This paper very cleverly illustrates a problem which can be highly difficult to explain without a concrete example. The repeated format of the experiment with changes between the discussion and conclusions really helps to hammer home the point. This paper will no doubt help ignite discussion on the topic of how unreliable p-values can be and how much of a factor chance can play.`  `The next time I'm asked by an undergraduate student (or a colleague) about the validity of p-values, I now know where to point them!`
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Derek Reilly - Dalhousie University, Halifax, Nova Scotia, Canada

Overall Rating	4 (Probably accept: I would argue for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	4 (Expert)
Conflict of Interest	Yes, I have a conflict

The Review	`-- review ---`
	`The authors introduce a methodological innovation that has the potential to revolutionize empirical research in HCI, and likely in other disciplines as well.`
	`Running studies across parallel universes opens up many intriguing`

	possibilities. This paper explores one of these: parallel universes as a mechanism for replicating studies. I am arguing for acceptance based on this very singular contribution. As with much groundbreaking work, however, there is one deep flaw that will need careful consideration in a final version of the paper. The authors state that they only approximated parallel universes by repeating a study sequentially in a single universe. I was surprised to find this out only quite later on in the paper. The authors should clearly state this as a limitation of their work from the outset -- if this is indeed the case. I have my suspicions, however, as parallel universes would very clearly explain the differences found in each replication. Do they have technology that they are unwilling or unable to share? -- review --- The authors introduce a methodological innovation that has the potential to revolutionize empirical research in HCI, and likely in other disciplines as well. Running studies across parallel universes opens up many intriguing possibilities. This paper explores one of these: parallel universes as a mechanism for replicating studies. I am arguing for acceptance based on this very singular contribution. As with much groundbreaking work, however, there is one deep flaw that will need careful consideration in a final version of the paper. There is also no discussion of the implications of alternate realities/universes on the particular study reported. In one universe humans may have webbed fingers with suction-cupped tips, or they may control sliders using laser beams emanating from their eyes. Multiple universes mock our scientific desire for replicability and generality. -- review --- The authors introduce a methodological innovation that has the potential to revolutionize empirical research in HCI, and likely in other disciplines as well. Running studies across parallel universes opens up many intriguing possibilities. This paper explores one of these: parallel universes as a mechanism for replicating studies. I am arguing for acceptance based on this very singular contribution. As with much groundbreaking work, however, there is one deep flaw that will need careful consideration in a final version of the paper. This reviewer wonders why they repeated the same study multiple times. Why not tweak one or two things each time -- that way you might find out five things, instead of one thing five times. -- review --- The authors introduce a methodological innovation that has the potential to revolutionize empirical research in HCI, and likely in other disciplines as well. Running studies across parallel universes opens up many intriguing possibilities. This paper explores one of these: parallel universes as a mechanism for replicating studies. Of course, the authors only approximated parallel universes by repeating a study sequentially in a single universe. In doing so they share some of the issues in common empirical approaches when interpreting results, by showing how different samples impacted the outcome of the study. -- review --- I don't see the science. Reject (0).
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Jeff Shrager - Symbolic Systems Program, Stanford, Stanford, California, United States

I'm sorry but I don't get it. This seems cute but silly. They didn't power their study right, and anyway a meta analysis would produce a pretty clear conclusion. If the point is that HCI people don't know how to design experiments, fine, but let's not confuse that with an indictment of statistics.

reply

Geoff Cumming - School of Psychological Science, La Trobe University, Melbourne, Victoria, Australia

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	4 (Expert)
Conflict of Interest	Yes, I have a conflict

The Review	This article is fabulous, in both senses! Legendary, in both senses! Out of this world, in about 9 senses! Many scientific disciplines either do not understand how seriously weird, deficient, and damaging is their reliance on null hypothesis significance testing (NHST), or they are in a state of denial. Over more than 50 years, leading scholars have explained cogently the deep flaws of NHST. A good review is by Rex Kline: http://tiny.cc/klinechap3 One dramatic failure of p values is that they are so spectacularly unreliable: Repeat an experiment, just the same but with a different sample, and you are very likely to get a very different p value. The variability of p occurs at any level of power, and is most spectacular at the middling levels of power typical of published research in CHI and many other disciplines. The article is an imaginative and striking presentation of the effects of this unreliability of p. If it seems too weird to be true, then investigate further--it is indeed fully justified. I said I have a "conflict of interest". Actually it's rather the reverse: No conflict, because I fully agree with the authors and their message. I'm the author of their refs [2] and [3]. I can mention a more recent dance of the means: http://tiny.cc/dancepvals2 I salute the dramatic example the authors have created, and hope many folks have a chuckle, then think deeply about the way they and their discipline does statistics. We don't need the security blanket of p, there are much better ways! Shortcomings of the article? I thought the message should have been pretty clear to anyone reading the last page. But some of the reactions and comments suggest some folks have missed the message. Should it be telegraphed even more clearly? One question to the 24 authors: How did they manage to achieve 8 people with the same name at the same institution, presumably drawing 8 salaries! Please tell us--we all want to join such an amazing gravy train! Geoff Cumming g.cumming@latrobe.edu.au La Trobe University, Melbourne, Australia
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Pierre Dragicevic - INRIA, Orsay, France

Thanks for all your comments!

We appreciate the many insightful suggestions. We will make sure to better examine the work of theoretical physicist Brian Green and his notions of quilted, cyclic and brane universes. For the moment we will try not to overgeneralize our results. There might be parallel universes where reliance on p values is not absurd at all.

As for the 9th universe where our doubles happily collaborate with Garth Shoemaker, we did not find a way to establish a reliable communication yet, but we are working hard on this issue.

We seem to be missing 3 reviews from Derek Reilly, and we hope nothing wrong happened to his 3 remaining doubles. We like to think they missed our paper because they were on vacation and/or enjoying their time. Please keep us updated.

Our paper can be interpreted in several different ways so perhaps we should clarify our intended message.

As amateur statisticians, our goal was modest. The purpose of our submission was only to raise awareness on the need to question our current practices (i.e, the way we typically use NHST - null hypothesis statistical testing). We are glad to see that our article seems to be achieving this goal.

We did not provide an extensive discussion on the numerous problems of NHST, nor did we discuss how statistics should be ideally done. This has been already discussed in hundreds of papers across decades in many related disciplines, and we thought there was no need for yet another paper on this.

At the end of our paper we refer our readers to the work of a statistics reformer who inspired our article, and whose arguments and recommendations we find especially simple and compelling. He wrote the review above. Make sure you check his "dance of the p values" and take a serious look at his work. An article summarizing his book is available here: http://tiny.cc/tnswhyhow

The point of Geoff Cumming's work (and of our paper) is NOT that we should use larger sample sizes. The idea that there is a "right" or "wrong" sample size is based on power analysis, which is based on NHST, which is based on an arbitrary alpha cutoff, and is therefore subject to the same dichotomous thinking fallacy. Sure the larger a sample, the better. The point is that no matter the sample size, we need to be more subtle in the way we look at the strength of evidence in our

data, and think (and communicate) in shades of gray. The estimation approach (computing effect sizes with confidence intervals) is the right "user interface" to do this.

Also, our point is NOT that we should reject statistics altogether. Confidence intervals are actually based on the very same statistical tools as NHST (t-tests, ANOVAs, etc.) and there is an equivalence between the two. It is just a very different way of presenting results, and it promotes a different way of thinking, as shown by studies on statistical cognition. NHST is not theoretically "wrong" or flawed, it tells us something real. But even when it is correctly used and interpreted, it does not tell us what we want to know, and it promotes a thinking style that perhaps is not the best suited to HCI.

Pierre, Pierre, Pierre, Pierre, Pierre, Pierre, Pierre, Pierre, Fanny, Fanny, Fanny, Fanny, Fanny, Fanny, Fanny, Fanny, Stéphane, Stéphane, Stéphane, Stéphane, Stéphane, Stéphane, Stéphane, and Stéphane.

reply

Jeff Shrager - Symbolic Systems Program, Stanford, Stanford, California, United States

Ah. I see what you are after, I think. But this is still a bit of an unsubtle way to go about making what is a subtle point, P-values and CIs are essentially the same thing (mathematically complementary), and are useful in different settings. If you're trying to get a sense of the range of ... well ... the effect ... then, sure, use CIs, but if, at the end of the day, you have to make a decision about whether to give someone a toxic, and potentially life saving chemotherapy (or foist a very difficult and expensive engineering problem on a team, or a very painful UI on a user community), that is, you have to make a decision, p-value are ... well ... valuable. I guess I stand by my claim that simply indicting p-values isn't the best way to approach teaching statistics. (BTW, psych stats are lame. Students should be forced to take biomedical stats where this sort of thing is front and center, and just second nature.)

reply

> **Jeff Shrager** - Symbolic Systems Program, Stanford, Stanford, California, United States
>
> Ps. We shouldn't be talking about any of this anyway. Closed ended experiments as so 20th Century. Everyone should be doing Bayesian adaptive trials. Look it up! :-)
>
> reply
>
> > **Pierre Dragicevic** - INRIA, Orsay, France
> >
> > Jeff: I agree that we should take into account the cumulative nature of knowledge rather than considering experiments in isolation. Geoff Cumming mentions meta-analysis as a possible approach (works very well with CIs), but this assumes we have replications (very uncommon at CHI for the moment). As for Bayesian statistics, I often see them mentioned as an alternative to NHST, but I have yet to find an introduction that's not overly technical and whose methods I can use in practice. I'd happy if you had any recommendation.
> >
> > reply
> >
> > > **Juan Sebastian Casallas** - Human-Computer Interaction Graduate Program, Iowa State University, Ames, IA, USA
> > >
> > > I really liked your article, in case you're interested in Bayesian statistics, I recommend "Doing Bayesian Data Analysis" by John K. Kruschke. It really provides a hands-on introduction to Bayesian statistics for "real people". http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/ (please don't judge the book by the "framed" website, or by the dogs in the cover :))
> > >
> > > reply

Pierre Dragicevic - INRIA, Orsay, France

Jeff, totally agree with you as far as clinical research is concerned. But when was the last time a decision maker had to choose between adopting interaction technique B or sticking to interaction technique A, with many lives at stake?

There may be exceptions but as a general rule, the way UI design ideas are adopted in consumer products is extremely different from the decision making process involved in health care, and for which we definitely need statistical tools that incorporate the notion of risk and allow us to do reliable cost-benefit analyses. Though some argue that NHST also does serious damage in this domain -- you are probably aware of the file drawer effect and have most likely seen Ben Goldacre's talks.

Most of our HCI design ideas are not adopted anyway, and doubt having $p < 0.0001$ is going to change this. So which decision exactly is helped by NHST? There is none. Yet I think studying new designs is important, as such studies contribute to our understanding of HCI. Certainly we are scientists and we need to rely on facts and update our beliefs based on weight of evidence. But it's hard for me -- and I bet for many HCI researchers who were introduced to NHST -- to see how concepts like Type I errors and alpha levels are any relevant in this context.

reply

> **Jeff Shrager** - Symbolic Systems Program, Stanford, Stanford, California, United States
>
> Great. So let's assert agreement on the statistical facts on the ground, and talk decision making. I agree that most of what one sees in CHI doesn't need the stats (or at least the sort of stats) that is generally offered. But most of the "important" UI work doesn't appear in CHI, and does involve either killing people or wasting huge amounts of money. I'd start talking military, but someone might start singing :-) so let's talk aviation (piloting or atc), medical technology (a huge an rapidly growing area of HCI), and just plain huge scale UI engineering at huge scale companies. Like, Apple and Google will spend god knows how much money changing something in the iPhone or droid, and could lose a zillion dollars in market share if they get it wrong. UI decision making really does matter.
>
> reply

Pierre Dragicevic - INRIA, Orsay, France

I was precisely thinking of aviation when I considered the possible exceptions. Although there are a few specialized venues, we don't see much of this work at CHI. Studies on medical technology: my opinion is that they shouldn't be published at CHI, but be judged by competent experts from clinical research. Apple and Google: I'd be curious to know whether or not they rely on NHST. Google publishes at CHI, Apple does not.

Yes UI decision making can matter a lot, but in the end very few studies published at CHI really involve that type of decision making. The ones who make a decision based on p values are really the paper's reviewers, because it makes their job easier and they were told it was important. This hampers the advancement of science for many reasons, one of which being the file drawer effect I previously mentioned.

reply

Jeff Shrager - Symbolic Systems Program, Stanford, Stanford, California, United States

Okay, I completely agree. Peace.

reply

Jeff Shrager - Symbolic Systems Program, Stanford, Stanford, California, United States

An aside, mostly in agreement with your point about reviewers, and said more to the community than to you personally (I know that you'll know most of what follows): The biomedical clinical research community requires that a competent statistician review all papers (at least in top tier journals) that have significant statistics, and "as a result" (it's not really as simple as that implies), almost all biomedical researchers now both employ competent experiment designers, and train their students (anyway those that will be involved in clinical research) in competent design and interpretation of experiments (or at least in knowing when they aren't competent themselves and have to get a consultation with someone who is). If they don't do this, they know they will not be able to get the result published (in a top tier journal). This doesn't 100% block trash from getting into the medical literature (esp. via the rapidly growing plethora of "open source" poser journals), but the general understanding of the importance and (some of the) subtleties and (some of the) problematics of experimental design and statistics in biomedicine...where it really does matter, and where they (mostly) really do think about these thing in great detail. In addition, as I mentioned above, there is a movement in clinical research to recognize what I called "bayesian adaptive trials", or, more generally, (bayesian) global cumulative analysis, wherein one can incorporate both subjective and objective data, and use this cumulative store of knowledge and data to make decisions when you need to make them. The GCTA approach tries to elegantly combine qualitative and quantitative "results", and if anyone is going to end up figuring it out, it's going to be biomed. So I'd suggest watching that space carefully.

reply

Jeff Shrager - Symbolic Systems Program, Stanford, Stanford, California, United States

Overall Rating

Appropriateness for alt.chi

Expertise

Conflict of Interest

The Review

```
This is not a review but my final reply to the deeply nested conversation
that blew out the conversational system...which is the funniest irony
I've seen possibly all year... Anyway, my final reply was supposed to
say: "I agree completely. Peace."

But this is a terrific commentary as well:

ERROR on line 11 of templates/showOpenComment.tmpl
  from line 57 of templates/showOpenComment.tmpl
  from line 59 of templates/showOpenComment.tmpl
  from line 57 of templates/showOpenComment.tmpl
  from line 59 of templates/showOpenComment.tmpl
  from line 57 of templates/showOpenComment.tmpl
  from line 59 of templates/showOpenComment.tmpl
  from line 57 of templates/showOpenComment.tmpl
  from line 59 of templates/showOpenComment.tmpl
  from line 787 of /templates/openReviewing.tmpl
  from line 823 of /templates/openReviewing.tmpl

Include files nested too deeply
```

Consider for commentaries

Reviewers nominations for commentary

reply

Anonymous

That's fixed. Sorry for the problem.

James Stewart
Precision Conference Solutions

reply

Committee Member 1

As alt.chi chair, I am proud that the discussion got so hot that you blew up the system.

reply

Chat Wacharamanotham - RWTH Aachen University, Aachen, NRW, Germany

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	4 (Expert)
Conflict of Interest	No, I do not have a conflict
The Review	Misinterpretation of p-values is a deep-rooted problem in both CHI and psychology. Comparisons of statistical results to illustrate the problem is a widespread instructional method from a classical literature, e.g., [A], to a recent CHI paper, e.g., [4]. Nevertheless, 8 x (Dragicevic et al.) unprecedentedly present a side-by-side comparison of associated interpretation of the results and the discussion in the context and style of CHI. For example, Universe #7 is a representative example of rationalizing non-significant results with anecdotal evidence which is unique to CHI community. Universe #2 & #6 and #3 & #8 pairs nicely contrast the abuse of the “highly” significant p-values in the discussion. Universe #4 and #5 contrasts the discussion of the interaction effects of different magnitudes. I believe that having illustrative examples tailored for CHI community will raise awareness of the danger of p-value reliance and increase the scientific rigorously of CHI. Besides, it is a great study material for HCI students and veteran alike. In summary, this paper contributes vivid examples of p-value problems in CHI context. This paper will add up to the attempt to improve qualitative analyses of CHI community by providing a concrete resource for discussion and education about the problem. Thus, I recommend accepting this paper. Suggestion: * One example of misinterpretation the degree of significance as an effect size would be adequate. Either #2 & #6 or #3 & #8 should suffice. * An exemplary analysis based on estimations (an alternative to do it right) would be useful for readers. * The generated data are in lognormal distributions, but all 8 universes seem to treat the data as if they were normally distributed. This violates an assumption of ANOVA. I surmise that the purpose of this omission is for the simplicity of the setup and the clarity of the p-value problem highlighted, but I’d recommend adding a sentence to acknowledge this limitation. * It is unclear what error bars in graphs represents. Should they represents SDs or CIs, this also contradicts the nature of the lognormal distribution resulting in a misleading graph. I’d recommend acknowledging this omission too. [A] Cohen, Jacob. "The earth is round (p<. 05)." American psychologist 49.12 (1994): 997.
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Pierre Dragicevic - INRIA, Orsay, France

Thanks for your great review and excellent suggestions.

The way you pick up the flaws in our 8 analyses is very pedagogical. We chose to keep discussions very minimal in our paper, as our goal was rather to provide illustrations that could spark discussions. For similar reasons, we did not provide an exemplary analysis, despite having received several requests for doing so. This could definitely be a great follow-up paper. Ideally, more than one. Since there no well-established ritual for interpreting results based on CIs alone, there is lots of room left for creativity and ingenuity.

The fact that we didn't log-transform our data did not escape your scrutiny. Good! This is quite common, and we wanted our analyses to be typical of HCI and authentic. The data was generated in R and independently analyzed in SPSS. I'm not sure whether it passed SPSS' normality tests, but with only 12 data points it seems possible.

Your suggestion of conflating p with effect size in some of the universes is nice, and we will consider improving our discussions. We already do this at several places (suffices to use the misleading word "significant") but it may be too subtle. We're definitely being too honest in Universe #1 when we acknowledge that the effect of technique is moderate.

reply

Per Ola Kristensson - University of St Andrews, St Andrews, Fife, United Kingdom

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)

Expertise	
Conflict of Interest	No, I do not have a conflict
The Review	This is a clever pedagogical demonstration of how a CHI researcher in "eight parallel universes" could plausible write eight different discussions depending on the outcomes of the p-values. Many papers, lectures and videos have talked about the problem itself before. However, I find this (to my knowledge) unique take by the authors to concretely spell out the different discussion sections that surely would arise in a CHI paper creative and insightful. I would like to use this paper when I teach HCI evaluation methods. A few remarks, which should not be taken as arguments for not accepting this excellent alt.chi submission: * As I believe another reviewer pointed out before, the data should have been log-transformed or analysed using a different model than an ANOVA. * I am not convinced the mathematical functions generate what I would call representative HCI data. It would have been excellent had the authors carried out one or two of the experiments with actual participants as well. * I am also not convinced the recommendation to use confidence intervals solves anything as confidence intervals and p-values are both tied to the same underlying frequentist models and thus share the same strengths and weaknesses. If the authors point is however that CHI papers should include effect sizes and error bars (confidence intervals) then I completely agree.
Consider for commentaries	
Reviewers nominations for commentary	

reply

Cosima Rughinis - Faculty of Sociology and Social Work, University of Bucharest, Bucharest, Bucharest, Romania

Overall Rating	5 (Definite accept: I would argue strongly for accepting this paper.)
Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	4 (Expert)
Conflict of Interest	No, I do not have a conflict
The Review	Writing a late review gives me the benefit of relying on previous ones. So, I can say that this is indeed a challenging paper - it is very visible in the debate it has sparked. I also like its writing concept, which nicely supports the argument and is also fun. Because of the multiverse approach, many reviews were also enjoyable reads, thus the benefit has amplified. For the sake of conversation, I think that (1) the paper touches an important issue, while (2) leaving some strongly related problems untouched, although it could have pointed towards them. (1) The important issue is over-reliance on statistical significance. This is indeed a dominant problem in quantitative analysis. At the very same time, there is already widespread criticism of this intellectual practice. I particularly like [1] and [2] but there are many others, and the authors also cite relevant work. So, in a way, this is an old combat zone already; the question rises whether the paper brings something new in this field. I think it does: the multiverse approach makes the argument more palatable (especially for students), so the party that opposes the use of statistical significance as proxy for substantive significance wins some ground. This is good news for me. (2) Still, the article itself highlights some other problems that remain untackled. Firstly, of all parallel versions of the researchers, only one group has thought to actually ask participants what it was like and to report their answers (in multiverse 7). While Chat Wacharamanotham reads this as "a representative example of rationalizing non-significant results with anecdotal evidence which is unique to CHI community", it is also the only place in the multiverse where participants actually get a voice. Secondly, all parallel researchers have opted to maintain the view of a universal human nature, reporting on 'users' that are assumed to share everything except some uninteresting random variation, maybe. The interesting variation is that of Technique and Difficulty - but what about differences within 'users'? And, moreover, who are these users, except for the fact that 2 are female? This says it all? This abstract participant, which is treated as implicitly representative for individuals in her/his kind across the world, without any trace of contextualization and distinction between types of users of uses is, I think, a strong limitation of experimental articles in the field of social and human sciences, including HCI. Therefore, I would suggest that at least some authors in some universes revise their reports to point to these two limitations, besides over-reliance on p-values: - An abstract, de-personified representation of participants / users; - A strict reliance on numbers at the expense of participants' meaningful feedback expressed with words.

	That is, let's multiverse more wildly. [1] S. T. Ziliak and D. N. McCloskey, The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor, MI: University of Michigan Press, 2008, p. 320. [2] C. Lambdin, "Significance tests as sorcery: Science is empirical--significance tests are not," Theory Psychol., vol. 22, no. 1, pp. 67–90, 2012.
Consider for commentaries	Yes, I would be willing to revise this review for inclusion in the paper's extended abstract.
Reviewers nominations for commentary	

reply

Committee Member 2

Appropriateness for alt.chi	3 (Appropriate - This paper is likely to promote debate in the CHI community.)
Expertise	2 (Passing Knowledge)
Conflict of Interest	No, I do not have a conflict
The Review	I think the reviews tell the story here. It's generating a lot of discussion already. That seems to me to be a winner for an alt.chi paper. It is interesting that the studies themselves found so much variance, we tend to assume that qualitative data is problematic in that it can't be replicated... this illustrates the same difficulty for quantitative data. Great. I would strongly encourage the authors to make good on Derek Reilly's suggestion about not leaving it until the end to explain the reveal. Also, if there's room to explain what a multiverse is. Not all of us have been reading there (and I had to go off and read about it). Indeed, I even wondered whether you needed that whole motivation. A simpler version is that you did the same study eight times and look there are differences.
Consider for commentaries	No, I do not want this review to be considered for commentaries.
Reviewers nominations for commentary	

reply

Add a comment