Proceedings of the ASME 2015 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
IDETC/CIE 2015
August 2-5, 2015, Boston, USA

DETC2015/DTM-46840

A SCALPEL NOT A SWORD: ON THE ROLE OF STATISTICAL TESTS IN DESIGN COGNITION

Mark Fuge
Department of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: fuge@umd.edu

ABSTRACT

The number of design studies using statistical testing has increased dramatically over the past decade. While this has benefits, statistical testing requires scrutiny to protect against common errors and misconceptions. To illuminate how these issues affect design, this paper provides a comprehensive analysis of the past decade of studies within the DTM community. Specifically, the paper 1) reviews the background of statistical testing across multiple fields, highlighting recommended practices, 2) discusses its use in the Design community, and 3) provides concrete methods for authors and reviewers to evaluate statistical tests employed in Design Cognition studies.

The analysis identifies recurring issues with: ignoring multiple comparisons; deficiencies in study and result reporting; inadequate defense of modeling assumptions; unavailable plots, data, and analysis files for replication; and lack of interpretation of statistical results with respect to practical outcomes or alternate forms of scientific inquiry. Based upon practices already adopted in other research communities, we put forth: 1) checklists that help authors and reviewers verify data reporting, analysis, and statistical assumptions; and 2) design guidelines for creating more reproducible design experiments. Ultimately, we argue that design researchers, reviewers, and editors should view statistical testing less like a sword and more like a scalpel—a specialized tool best used in concert with other techniques—to gain a more complete picture of Design Cognition.

INTRODUCTION

Engineering Design is complex, and design researchers choose among many kinds of scientific inquiry to investigate its physical and social phenomena. One type of inquiry has become increasingly popularly over the past half-decade: Null-Hypothesis Statistical Testing (NHST), or the use of statistical inference techniques to draw causal conclusions from data. (DTM studies incorporating NHST have risen from 9% in 2003 to 50% in 2014—Fig. 1.) Its increased use coincides with a rise in Design Cognition and Design Behavior research [1]: studies that borrow techniques from psychology to help us understand phenomena ranging from analogical reasoning [2] to creativity [3] to prototyping behavior [4, 5]. Our field has gained innumerable benefits from these advances, opening up many fruitful avenues for understanding design cognition.

However, researchers should approach any wide-spread adoption of particular techniques with careful consideration and sufficient understanding: What strengths and weaknesses does one approach to scientific inquiry have over another? How can we adopt best practices while ameliorating any known downsides? What can the history of these techniques tell us about pitfalls or opportunities for advances?

This paper offers some history and perspective on NHST techniques, reviews the usage of NHST techniques in Design, and highlights how the use of NHST techniques does not follow best practices encouraged by the statistical community. It initiates a critical conversation around research practice within the DTM community in the same spirit as others have done for

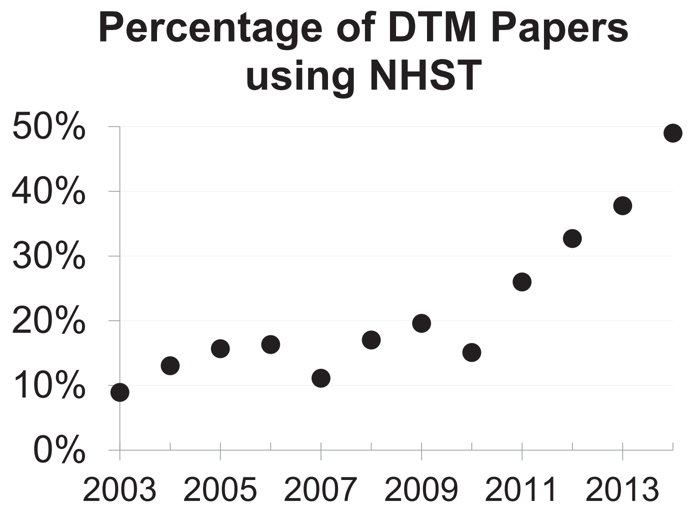

FIGURE 1. The percentage of studies using NHST has increased substantially over the past half-decade, to almost 50% of papers at the 2014 DTM conference.

Psychology [6], Medicine [7], Management [8], and Human-Computer Interaction [9].

The paper first summarizes the NHST debate across multiple fields. Then it reviews the papers from the past decade of the International Design Engineering Technical Conference's Design Theory and Methodology track. The review answers the following questions:

1. How has DTM's use of NHST changed over time?
2. To what extent do statistical issues present in other fields occur within design experiments?
3. How can design researchers avoid repeating other fields' mistakes in the future?

The paper then offers up some guidelines to make it easier for design researchers, reviewers, and editors to evaluate statistical testing and prevent issues rampant in other fields from propagating to Design Cognition (and Design more broadly).

RELATED WORK

Throughout the history of NHST, but particularly over the past two decades, researchers have debated what role NHST should play in scientific inquiry. These debates fall roughly into three major categories, which we order by approximate scope: 1) how should one correctly execute a particular statistical method, 2) how should one use statistics to support an argument, 3) what role should statistical testing play with respect to other forms of scientific inquiry? We refer to these three types of debates as *execution* issues, *design* issues, and *interpretation* issues, mirroring the divisions used by [10] and [11].

Execution issues concern whether or not a particular statistical method is justified, given the data and study design. Examples include concerns over whether data match assumptions that a method uses; *e.g.*, using tests which assume normality when the data are non-normal [12]. Or whether a given sample size suffices to claim results of a particular strength at a particular power [13–15]. Or how researchers interpret the a statistical result (such as the conflation of NHST as proposed by Fisher versus Neyman-Pearson [16, 17]).

Design issues concern whether the study design and statistical model are appropriate (regardless of the methods used to calculate the outcome). Examples include using linear or logistic regression models in observational data, when the data are neither linear or exogenous [18]. Or the Multiple Testing Problem—simultaneously using multiple statistical tests and not adjusting the threshold (α) to compensate for the increased false positive rate—causing spurious statistical results [19, 20]. Or selective result reporting that makes it difficult for others to compare across studies (*e.g.*, for effect size) and hide researcher degrees-of-freedom (such as only selectively reporting experimental conditions), biasing the results [6].

Both Execution and Design issues result from a researcher's methodological errors. It has led to calls for greater statistical literacy [21, 22] and changes in professional guidelines [23, 24]. Design issues are more systemic, requiring fundamental changes in the statistical models a community uses. For example, researchers still use linear regression models on non-randomized data, despite clear evidence (including mathematical proofs) that those models can be misleading [12, 18, 25]. Research communities have pushed towards increasing replicability of experiments by publishing data [26, 27], instituting editorial checklists [22, 28, 29], pre-registering experiments [30, 31], and even independently replicating common experimental results [32–34].

Debates extend beyond methodological arguments to philosophical questions about the role of NHST in the scientific process. These debates come in two forms: outcomes and methods.

For outcomes, the central question has been "If an outcome is statistically significant, does that mean it is practically significant?" [10, 11, 35] NHST opponents argue that measures like Effect Size are more useful than rejection of a null hypothesis [10, 35], or that the null models used are not realistic enough to be useful [8]. NHST proponents argue that statistical significance *can* imply practical significance, depending on the goal of the research, and that thoughtfully performed testing with appropriate models is an indispensable scientific tool [11, 36].

For methods, the central question has been "What is the most appropriate mechanism for establishing truth and forward scientific progress?" Essentially, researchers and philosophers have debated the relative merits of the Hypothetico-Deductive model (of which NHST is a corner-stone) compared to other models, such as deductive or abductive methods [37–39].

This relates to Design in that we use many quantitative and qualitative techniques that correspond to different scientific methods. This paper views this diversity as one of the strengths of our field. Researchers have sought to improve design research practice, such as incorporation of placebo controls within design studies [40], comparing lab studies with project and industry practice [41,42], and enumerating validation strategies for design models [43]. However, none of those have looked specifically at NHST's role within Design Cognition.

METHODOLOGY

This paper's corpus consists of all articles presented in the Design Theory and Methodology Conference from years 2003 to 2014, totaling 603 articles. Many articles do not perform any statistical testing. To identify just those articles that use statistical testing, we used the following procedure:

1. An automated script analyzed each article's text for any words that matched the following case-insensitive patterns: "statistic*" or "signific*". Of the 603 articles, 493 matched those patterns, and the matched sentences were output to a text file.
2. The author manually skimmed the extracted sentences from each of the 493 articles to determine, in the context of the sentence, whether the sentences may have been referring to a statistical test or a conclusion from such a test. If the extracted sentences did not contain sufficient information to make that judgement, we read the entire article to verify whether it used any statistical tests.

We identified 130 articles out of 603 ($\approx 22\%$) that used some form of statistical test. From these, we reviewed two-thirds (≈ 88) in detail by reading the methodology, results, and discussion sections of the articles, and annotating each article with codes from the three major categories mentioned above: Execution issues, Design issues, and Interpretation issues. The codes were derived from existing reporting checklists or guidelines [22,28,29,44–46].

Execution Issues

These codes address execution of specific statistical models:

Sample Size: Coded "Yes" if: the study mentions how the sample size was determined. Coded "No" if: otherwise.

Power: (*i.e.*, *sensitivity* or *type-II error*) Measures a test's ability to detect an effect of a particular size when one actually exists. Coded "Yes" if: the text mentions the intended power of the study. Coded "No" if: the text does not mention the intended power.

Effect Size: Coded "Yes" if: text clearly states the effect size in either absolute or relative (*e.g.*, Cohen's d) terms. Coded "No" if: otherwise.

R^2: (*i.e.*, *Explained Variance*) Measures a statistical model's goodness-of-fit to data. Coded "Yes" if: text provides R^2 values for any relevant statistical models in the article. Coded "No" if: otherwise. Coded "N/A" if: the tests used do not have a straightforward interpretation of explained variance.

Tests Assumptions: Coded "Yes" if: the paper attempts to verify any assumptions underlying any of the statistical tests used (*e.g.*, normality, heteroscedasticity, *etc.*). Coded "No" if: otherwise.

Design Issues

These codes address the design, choice, reporting, or interpretation of statistical models given experiment data:

Intention-to-treat (ITT): Coded "Yes" if: any text mentioned whether participants dropped out of the study. Coded "No" if: otherwise. Coded "N/A" if: dropout was irrelevant to the study design.

Exclusion Criteria: Coded "Yes" if: text describes whether any data were discarded when calculating the final statistics (or states that no data were excluded). Coded "No" if: text did not state whether it excluded any data.

Multiple Comparisons: Coded "Yes" if: multiple tests were performed and any were adjusted to account for the increased false positive rate. Coded "No" if: multiple tests were performed, but no adjustment was conducted or mentioned. Coded "N/A" if: only a single test was conducted.

Data Plots: Coded "Yes" if: the article provides data plots for at least one statistical test. Coded "No" if: otherwise.

Accessible Data: Coded "Yes" if: article provides a link to data used in statistical tests, or describes why such data cannot be provided (*e.g.*, medical records). Coded "No" if: otherwise.

Accessible Analysis or Code: Coded "Yes" if: article provides a link to data analysis code, or describes why such code cannot be provided (*e.g.*, non-disclosure agreements). Coded "No" if: otherwise.

Interpretation Issues

These codes address the interpretation of the study:

Interprets Magnitude Outcomes of Effects: Coded "Yes" if: article attempts to put the NHST results into a real-world context (*e.g.*, talking about the practical magnitude of the size of an observed effect in real-world terms). Coded "No" if: only "Significant or not Significant" interpretations were provided.

Included Alternate Forms of Inquiry: Coded "Yes" if: article presents any data or interpretations beyond NHST, by themselves or used in conjunction to explain the quantitative results. Coded "No" if: only NHST results were present.

TABLE 1. Percentage of Studies with Given Code Labels. Higher numbers in "Yes" or "N/A" are better than those in "No."

Category	*Annotation*	*Yes*	*No*	*N/A*
Execution	Sample Size	7	93	–
	Power	2	97	1
	Effect Size	6	94	–
	R^2	21	71	8
	Tests Assumptions	8	86	7
Design	Intention-to-treat	19	69	11
	Exclusion Criteria	40	37	23
	Multiple Comparisons	7	86	7
	Data Plots	79	21	–
	Accessible Data	1	99	–
	Accessible Analysis	1	99	–
Interpret.	Outcome Significance	22	78	–
	Alt. Forms of Inquiry	22	78	–

RESULTS AND DISCUSSION

We divide our analysis into three parts: 1) annotation codes across the corpus, 2) distribution of statistical tests used, and 3) qualitative excerpts from individual papers in the corpus that reveal representative attitudes about statistical tests in Design Cognition studies. We have made all data and analysis files available on the paper's companion website, for those who wish to replicate our results.[1]

Summary of Annotation Codes

Using the coding scheme defined in the above section, Table 1 lists the code prevalence across the corpus. We designed the codes so that "Yes" or "N/A" codes were preferred to "No."

Execution Issues

Sample Size: Most papers (93%) did not mention how they selected their sample size or stopping rules for collecting data. The American Psychological Association (APA) recommends including this information [45, Table 1].

Power: Most papers (97%) did not mention the intended Power of the employed statistical tests. Of those that did, most calculated the power of the test after the observed effect size, which over-estimates the actual power of the test [13, 15]. The APA encourages disclosing power calculations for study design and sample size calculations prior to data collection [45, Table 1].

Effect Size: The vast majority of papers did not mention the expected or observed Effect Size for their primary outcomes, instead listing only "Significant" or "Not Significant" results. The APA encourages reporting both test statistics (t/p-values) as well as effect size measures and power calculations, since researchers need that information to compare effects across studies [45, Table 1].

$\mathbf{R^2}$: For studies that fit statistical models to data (*e.g.*, regression models), 86% failed to report the variance explained by the model. Readers need this information to determine whether a model is appropriate, since p-values, power, and effect size alone do not provide that information.

Tests Assumptions: Few papers verified any of the assumptions required by their statistical tests. Certain tests are more sensitive to model assumptions violations than others (*e.g.*, F-Tests, regressions [12, 18]). A small number of papers used model diagnostic tools (*e.g.*, Normality tests, equal variance tests, *etc.*). Generally, such diagnostic tests can fail to reject inappropriate models with sufficiently high power [25, 47], so statisticians recommend providing data plots for additional verification [12, 48].

Design Issues

Intention-to-treat (ITT): Most papers (69%) did not mention whether or not participants dropped out or switched treatment groups during the study. The APA encourages reporting participant drop-out or switching, even if none occurs [45, Table 1]. In cases of dropout, certain types of statistical models need to be adjusted [49].

Exclusion Criteria: Certain papers were structured such that participant data might be excluded from the final analysis (*e.g.*, outliers, lack of participant reporting, unreadable data, *etc.*). Of these, roughly half did not describe the criteria by which data was excluded or mention whether data was excluded. Since certain statistical tests are biased by data exclusion [49], the APA encourages authors to report exclusion criteria, even if no data is excluded [45, Tables 1&4].

Multiple Comparisons: Most articles used less than 20 simultaneous tests, however several utilized over 100 simultaneous tests without correcting for the (substantially) increased False Positive rate [6,19,20]. Many journals now require adjusting for multiple comparisons [19,50] prior to publishing an article (*e.g.*, Nature [28]).

Data Plots: Most papers provided at least one visual plot of the data. This is important for determining appropriate statistical methods [12,48].

Accessible Data: Only one paper in the corpus provided access to the data required to verify the authors' results. This stands

[1] `http://ideal.umd.edu/dtm_stats`. We anonimized the specific paper titles and authorship as a courtesy to other authors. We can provide decoding keys upon request to authors who wish to conduct a full replication.

in contrast to calls from government agencies [51] and journals [27, 46, 52] for the availability of (anonimized) data to increase research transparency.

Accessible Analysis or Code: Only one paper in the corpus provided access to the computer programs or analysis files required to verify the authors' results. This stands in contrast to calls from journals for replicability and transparency in scientific reporting [27].

Philosophical Issues

Interprets Magnitude Outcomes of Effects: Most papers (78%) did not interpret the magnitude of their statistical tests in terms of practical effects, preferring to only state whether an effect was "Statistically Significant" or not. While it is true that laboratory effects will differ from those seen in real-world practice, we view this oversight as a lost opportunity for a more thorough interpretation of whether an intervention or behavior is worth investigating further.

Included Alternate Forms of Inquiry: Most papers (78%) did not perform alternative forms of scientific inquiry beyond analyzing NHST results. Those that did typically provided one or more of the following: 1) qualitative interview excerpts from participants, 2) specific case studies on one or more data points, or 3) refinement and interpretation of a computational model separate from those used to generate the NHST results.

Distribution of Statistical Tests

Table 2 summarizes the distribution of statistical tests used in each study. The counts are not mutually exclusive within studies: if a study used both t-tests and linear regression we added a count to each entry. The most popular statistical tests were the T-Test (32), One-Way ANOVA (26), Pearson Correlation Coefficient (29), and various Generalized Linear Model Regressions (GLR) (20). In general, researchers used T-Tests and ANOVA when comparing new interventions (*e.g.*, design methods) in laboratory settings, while they used Correlations and Regressions for observational data (*e.g.*, classroom or historical data studies). That said, there were a large variety of study designs outside of those two cases, and a full review of articles' content is outside the scope of this paper (See [1] for an overview).

Given the prevalence of the above tests, it is useful to understand some of the assumptions underlying these tests, as per Freedman [12, pg. 101]:

> *"Estimation and significance testing require statistical assumptions. Therefore, you need to think about the assumptions—both causal and statistical—behind the models. If the assumptions dont hold, the conclusions dont follow from the statistics."*

Regardless of whether the goal is to identify causation or correlation, researchers need to understand the underlying mechanics of the models they use and when they might be in error. Below we highlight key notes about the application of each of these tests and point out which assumptions tend to be robust to violations.

Type	*Test*	*# Studies*
Location	T-test	32
	Proportions t-test	3
Category	Pearson's Chi Squared	7
	Fishers Exact Test	1
ANOVA	F-test/one-way	26
	two-way	3
	MANOVA	1
	ANCOVA	2
Regressions	Linear Regression Coeff.	9
	Ordinal Logistic	4
	Stepwise Regression	2
	Probit Regression	1
	Logistic Regression	4
Correlation	P. Correlation Coeff.	19
	Spearman Rank Correlation Coeff.	8
Non-Parametric	Kruskall-Wallis	6
	Conf. Intervals	5
	Permutation test	1
	Mann-Whitney U / Rank Sum Test	9
	Kolmogorov-Smirnov (K-S)	1
Inter-rater	Cronbach's Alpha	4
	Pearson Correlation	10
	% Agreement	1
	Intraclass correlation	2
	Krippendorff's alpha	2
	Fleiss Kappa	1
	Cohen's Kappa	12
Other	Cohen's D	1
	F-Test Variance	1
	Bionomial-CDF	1
	Levene Test	3
	Logit	1
	Factor Analysis	1
	Shapiro-Wilk test	3
	Mauchlys test	1
	Fisher's Least Sig. Diff.	2
	Unclear Test	2

TABLE 2. The most common tests used in the study corpus were: T-Tests (32), One-Way ANOVA (26), Pearson Correlation Coefficient (29), and various Generalized Linear Model Regressions (20).

T-Test, ANOVA F-Test, and Generalized Linear Regression The T-Test is one of the most wide-spread hypothesis tests for differences in mean of two variables. ANOVA (and family) are multi-variate extensions of it. They all assume each observation is the sum of causal effects plus-minus some errors that are independent of each other, are identically distributed around 0, and have finite variance. For n data points with p variables ($p = 1$ for t-test, $p > 1$ for ANOVA and GLR), $n - p$ should be large (Central Limit Theorem) or the errors in the test statistic should be normally distributed [12, pg. 70]. If the errors are not distributed around 0, then the estimates will be biased. If the errors are distributed around 0, but they are not independent of each other or the treatment levels then the estimator will be unbiased, but formulas for the standard error (and therefore p-values) will be inaccurate [12, pg. 68].

There are three common situations in design experiments where the above assumptions would causes issues. First, when the number of participants per treatment group is small, assumptions about normality of errors would be difficult to verify, even with diagnostic tests [47]. In these cases either run studies with more participants or use tests that do not rely on the normality assumption.

Second, when the treatment conditions are not randomized, assumptions regarding treatment-error independence (*exogeniety*) do not hold, causing misleading standard error estimates. This is common in design cognition studies based on observations of in-class or semester-long activities (or any other case where participants have not be randomly allocated to conditions). In these cases, ANOVA, Regression Coefficients, and the F-Test can mistakenly "reject" a null hypothesis based on inaccurate standard error formulas. Without randomization (or some other means of establishing exogeniety, such as Instrumental Variables [12, pg. 181]), one cannot draw causal conclusions from observational data using statistical testing alone.

Third, even when assumptions regarding normality, *etc.* are reasonable, many papers draw conclusions about particular aspects or coefficients of a model by arguing that the omni-bus F-test statistic for the entire model is significant (rather than breaking down the model into simpler effects, contrasts, or post-hoc tests controlled for multiple comparisons on particular coefficients). Freedman [12, Ch. 5.7] succinctly highlights why this might be a problem:

> *"If F is significant, that is often thought to validate the model. Mistake. The F-test takes the model as given. Significance only means this: if the model is right and the coefficients are 0, it was very unlikely to get such a big F-statistic. Logically, there are three possibilities on the table. (i) An unlikely event occurred. (ii) Or the model is right and some of the coefficients differ from 0. (iii) Or the model is wrong. So?"*

Pearson Correlation Coefficient NHST tests whether a correlation coefficient is non-zero. The Pearson Product-Moment Correlation Coefficient assumes linear dependence between two variables and that the data has constant variance (*homoscedasticity*); when dependence is non-linear or heteroscedastic, the correlation coefficient will be inaccurate [48,53]. When estimating the probability that a coefficient is non-zero, researchers can employ several tests, each with its own assumptions; in our corpus most papers did not report the specific method used to estimate the desired probability. Common options include permutation tests, which assume exchangability, or t-tests and Exact tests which assume normality. The Pearson Correlation Coefficient is sensitive to data containing outliers [53].

Qualitative Excerpts

Several papers made specific statements that highlighted some cultural reasons for why NHST has been so rapidly adopted in our field (Fig. 1). We want to discuss the rationale behind some of these statements and dispel some common misconceptions regarding NHST. These break into three main categories: using causal language to describe correlational observations, confusing what a rejected statistical test means, and using statistical testing as substitute for other forms of scientific inquiry.

We found common confusion between correlation and causation, which typically occurred in a two-part pattern. First, the result section would present graphs, correlations, regressions, *etc.* that described how certain behaviors correlated with outcome measures (*e.g.*, grades, creativity scores, self-assessment, *etc.*). Next, the discussion and conclusions section would interpret those correlations as causal; for example through statements such as "Our results demonstrate that designers should do X to achieve Y" or "Our results support Z's research that X increases Y". The papers would not usually validate the model's causality prior to making this tenuous leap of faith from correlation to causation. While one might be tempted to give this gap in logic the benefit of the doubt, it creates a problem for our community, since, once published, false claims tend to persist in literature despite copious subsequent contradictory evidence [54].

A second common confusion stemmed from what statistical tests allowed one to claim. Several papers stated that because NHST rejected the null, the intervention was "statistically proven" to be effective; this common misconception is incorrect [17]. Likewise absence of evidence was taken as evidence of absence:

> *"Interestingly, however, analysis showed that there was no statistically significant correlation between* *removed for anonymity* *and the scores that the teams received in their evaluation. In other words, there seemed to be no grading bias based on* *removed for anonymity*."

Generally, failing to reject a result does not indicate no effect; it more commonly indicates a lack of power in the test [15].

Lastly, certain papers used NHST as a substitute for other forms of scientific inquiry:

> *"Unlike many studies of actual design processes, we use powerful statistical analysis tools to gain insight into the data rather than qualitative, case-based techniques."*

We view this mutual exclusion as counter-productive. While statistical techniques are indeed powerful (when used correctly), one needs to balance that power with an appropriate degree of control and caution. Rather, it is in our best interest to combine multiple forms of inquiry where possible, recognizing the limitations of each [55].

Implications on Results of Published Articles

Given the statistical issues noted above, what does this imply for the knowledge contained in existing articles? Depending on the particular case, this ranges from provoking inconvenience to invalidating certain results. Of the 13 issues shown in Table 1, three broad categories could have different impacts on each article's contributions.

First, unavailability of data plots, raw data, and analysis files creates inconvenience for readers and researchers, but does not affect the accuracy of knowledge contributed by an article (assuming the authors did not selectively present or filter the data used in their article). We should strive for transparency and more useful reporting in the future, but past articles raise no immediate concern.

Second, issues relating to Effect Size, Power, measures of outcome significance, and alternative forms of inquiry, make it harder for readers to evaluate the external validity of the knowledge. This does not imply incorrectness, but rather difficultly in correctly interpreting a study beyond just "statistically significant or not," reducing the article's utility.

Third, issues relating to sample size, R^2, testing assumptions, multiple comparisons, intention-to-treat, and exclusion criteria can raise questions regarding an article's internal validity (*i.e.*, a result's accuracy). In these cases, results might overestimate effects or, in the worst case, be plainly inaccurate. These types of issues could cast serious doubt on the knowledge contained in an article.

RECOMMENDATIONS FOR DESIGN RESEARCH

The above results indicate several areas specific to Design Research where we might focus our efforts as a community:

Pair statistical results with additional evidence. Certain design research studies cannot use randomization or other techniques to isolate causal behavior, for example, 1) studies conducted in classroom or workplace environments where researchers cannot randomize interventions (*e.g.*, new design methods or practices) to blind conditions; or 2) studies analyzing past design data where researchers performed no direct intervention (*e.g.*, regressions on company or team performance, such as stocks, design awards, grades, or behavioral patterns). In these cases, relying solely on evidence from statistical models will be misleading, since many assumptions (exogenity, independence, *etc.*) may not hold and will be difficult to verify. In these cases, we recommend conducting additional forms of data collection (*e.g.*, qualitative analysis or interviews, critical reviews of established theoretical models, supporting computational models, *etc.*) that can build a diverse portfolio of evidence for the desired phenomenon.

Match statistical assumptions to study designs and limit conclusions where appropriate. Certain study designs require careful analysis and interpretation: Were participants randomized to conditions? Could they stop using a particular design method (drop-out), or switch to a different one than they were assigned (cross-over)? Could participants do something that would disqualify their results (*e.g.*, leave a workshop, drop a class)? Are you testing multiple different outcomes simultaneously (*e.g.*, Novelty, Variety, Quantity, and Quality of ideas produced) or across a range of time periods (*e.g.*, multiple weeks during a semester, multiple time windows in a 30-minute session, *etc.*)?

All of those (and more) can limit the type of statistical models one can use and how one is allowed to interpret the results. In some cases, the data can identify causal behavior if corrected using appropriate techniques (*e.g.*, in Multiple Testing, or in drop-out), and in other cases the data might be drastically misleading (*e.g.*, in cross-over, or when analyzing past observational data that was not properly randomized). In either case, design researchers carry the burden of describing where and how our statistical models might be in error.

Standardize or better explain design outcome measures. To permit statistical analysis, design researchers quantify desired outcomes (*e.g.*, ideation, behavioral patterns, participant self-assessments, educational outcomes, company performance *etc.*) through a variety of measures (*e.g.*, creativity metrics, task or protocol frequency, Likert-style surveys, grade distributions on assignments or courses, biological signals such as fMRI and EEG, and stock performance, among others). This variety makes design interesting and diverse. However, it also requires careful interpretation about how (and by how much) interventions affect designers, what those changes mean, and to what extent "statistically significant" might not be "practically significant".

For example, if one created a new design method, tested that method on randomized participants, and recorded a statistically significant difference in the variety of the ideas generated, one would still have to ask "where did this significance come from (low within-sample variance, or large difference in means), and what does this difference mean in practical terms?" If the difference was $+0.4$ points on some chosen variety metric, what does that change represent in real-world terms? Is it worth the effort required to implement the method? Does a $+0.4$ increase mean the same thing at different locations on the measurement scale (*e.g.*, on Likert-scales from 0-10 points)? These are important questions that help us connect the statistical results to practical ones. While purely statistically significant results might be all that we are after in some cases, the burden still rests upon the researcher to discuss the utility that their results might achieve in practice.

Increase ease of replication by sharing data and study designs
At present, most reported design studies do not publish their data or study designs for future analysis. This makes it difficult for others to analyze past data (*e.g.*, for meta-analysis studies), to compare one-self with prior work, or accurately replicate previous effects. Fields such as Computer Vision, Machine Learning, and Network Analysis have benefited immensely from common experimental datasets and procedures shared by the community. Some efforts have been pursued to do the same for design (*e.g.*, ASU's Design Protocol Repository [56], Oregon State's Design Repository [57], and in [58–60]), however our field and review process has yet to establish the expectations and culture around replication and sharing seen in other data-intensive fields.

CONCLUSIONS AND RECOMMENDATIONS

This paper explored how Null Hypothesis Statistical Testing (NHST) has been used in Design Cognition studies, specifically, those in the Design Theory and Methodology conference. It did this via annotating articles using codes commonly used in review checklists and guidelines from other fields that apply NHST [22, 28, 29, 44–46]. It reviewed the most common statistical modeling assumptions, drawing on scenarios in design research where those assumptions will likely be violated.

From this, it found several issues that our community needs to address to conform with recommended statistical practices: ignoring multiple comparisons; deficiencies in study and result reporting; inadequate defense of modeling assumptions; unavailable plots, data, and analysis files for replication; and lack of interpretation of NHST results with respect to practical outcomes or alternate forms of scientific inquiry. Despite this, we do believe that NHST is useful to the scientific process, when conducted correctly, and we encourage researchers to familiarize themselves with the advantages and limitations of NHST. We hope this article can start a conversation within our community about collecting an appropriate and diverse toolbox of scientific methods for studying Design; one that builds on our multidisciplinary strengths without adopting the historical mistakes of those who came before us. One area for future research would be to followup on subsequently published journal articles and to compare the reporting standards between the two; one might expect the journal review process to ameliorate some of these issues.

In closing, we recommend the following actions for authors, reviewers, and editors to improve NHST use in Design Cognition studies and Design more broadly. These were built upon guidelines recommended by others [22, 23, 26, 29, 37, 61].

For Authors

Encourage Statistical Literacy: Encourage fellow researchers to take a graduate-level course in applied statistics offered by a statistics department. Consult graduate-level textbooks [12, 25] and modern reporting guidelines, such as the APA [45]. Many universities also offer free Statistical Consulting.

Use a Review Checklist: When planning and writing your research, use one of many published reporting checklists (*e.g.*, Nature's [28]) to catch common issues before submission.

Separate Out Confirmatory from Exploratory Testing: If you observe a set of possibly interesting effects, use a direct, targeted replication study to confirm an effect.

Consider Study Pre-Registration: If you *do* plan on conducting a mix of confirmatory and exploratory tests with one experiment, consider pre-registering [31] your study to avoid multiple testing or researcher degrees-of-freedom issues [6].

Consider Combining Forms of Inquiry: Conducting additional forms of inquiry, such as case-studies, qualitative analysis, or computational modeling can complement your NHST results.

Encourage Comparisons or Replications Across Contexts: Whether across labs at different universities, or across academic versus industrial environments, running parallel studies helps provide useful insight about robustness of design interventions (*e.g.*, Hernandez *et al.* [41]).

Encourage Transparency and Reproducibility: Where possible, make your analysis procedures and data available to others for review and later analysis. It encourages transparency, helps in meta-analysis studies or others types of reviews, and improves the impact of your articles [26, 27, 52].

Provide Graphs of Critical Data: Plotting data provides more insight than simply enumerating statistical properties [48]. Often many concerns raised during the review process can be avoided through proper display of primary data.

For Reviewers

Use a Review Checklist: When reviewing articles, use one of many published reviewing checklists (*e.g.*, Nature's [28], APA's [45]) to ensure authors provide recommended experimental detail in accordance with best practices.

Reviewer Statements: If you are concerned about possible backlash or unblinding of the review process, there are several standardized reviewer statements that can be included in all of your reviews (*e.g.*, the Center for Open Science's "Standard Reviewer Statement for Disclosure of Sample, Conditions, Measures, and Exclusions"[2]).

Encourage Transparency and Reproducibility: Request that analysis procedures and data be made available for review and later analysis. This is becoming common practice in many journals (*e.g.*, Nature [26, 27]).

Request Graphs of Critical Data: For critical statistical tests, request illustrative plots of that data, instead of just enumeration of statistical properties [48]. This helps future readers verify some common assumptions.

Emphasize Additional Forms of Scientific Inquiry: Using NHST as the sole arbiter of truth ignores many other potential forms of scientific inquiry. Encourage authors to expand on NHST results with other forms of inquiry to add depth to numerical results.

Shift the Discussion to Alternate Measures: Encourage discussion not just around "significant or not significant" (and at what level) but rather around the sizes of observed effects and their practical implications.

For Editors and Associate Editors

Institute Review Checklists for NHST Reporting: Many other journals provide a review checklist for papers involving NHST [28, 45]. These remind reviewers to check for NHST pitfalls. Some journals even require pre-registration of important experimental trials [30, 31].

Encouraging Responsible Data Sharing: Many journals encourage sharing of experimental data and code, subject to appropriate IRB and privacy restrictions [46, 51, 52]. This increases the research transparency and citation count (and thus impact) of individual articles and the journal.

Equip Your Journals to Accommodate Non-Print Media: If your journal does not yet allow so, encourage a mechanism for storing and indexing data or code associated with published articles (*e.g.*, a supplemental data submission procedure).

REFERENCES

[1] Dinar, M., Shah, J. J., Cagan, J., Leifer, L., Linsey, J., Smith, S. M., and Hernandez, N. V., 2014. "Empirical studies of designer thinking: Past, present, and future". *Journal of Mechanical Design,* ***137***(2), Nov., pp. 021101+.

[2] Hey, J., Linsey, J., Agogino, A. M., and Wood, K. L., 2008. "Analogies and metaphors in creative design". *International Journal of Engineering Education,* ***24***(2), p. 283.

[3] Yang, M., 2009. "Observations on concept generation and sketching in engineering design". *Research in Engineering Design,* ***20***(1), pp. 1–11.

[4] Häggman, A., Honda, T., and Yang, M. C., 2013. "The influence of timing in exploratory prototyping and other activities in design projects". In ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers.

[5] Toh, C. A., Miller, S. R., and Okudan Kremer, G. E., 2014. "The impact of Team-Based product dissection on design novelty". *Journal of Mechanical Design,* ***136***(4), Jan., pp. 041004+.

[6] Simmons, J. P., Nelson, L. D., and Simonsohn, U., 2011. "False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant". *Psychological science,* ***22***(11), pp. 1359–1366.

[7] Ioannidis, J. P., 2005. "Why most published research findings are false.". *PLoS medicine,* ***2***(8), Aug., pp. e124+.

[8] Schwab, A., Abrahamson, E., Starbuck, W. H., and Fidler, F., 2011. "Perspectiveresearchers should make thoughtful assessments instead of null-hypothesis significance tests". *Organization Science,* ***22***(4), pp. 1105–1120.

[9] Cairns, P., 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, BCS-HCI '07, British Computer Society, pp. 195–201.

[10] Ziliak, S. T., and McCloskey, D. N., 2004. "Size matters: the standard error of regressions in the american economic review". *The Journal of Socio-Economics,* ***33***(5), pp. 527 – 546. Statistical Significance.

[11] Lunt, P., 2004. "The significance of the significance test controversy: comments on size matters". *The Journal of Socio-Economics,* ***33***(5), pp. 559–564.

[12] Freedman, D., 2009. *Statistical models: theory and practice*. Cambridge University Press.

[13] Hoenig, J. M., and Heisey, D. M., 2001. "The abuse of power". *The American Statistician,* ***55***(1), Feb., pp. 19–24.

[14] Ingre, M., 2013. "Why small low-powered studies are worse than large high-powered studies and how to protect against trivial findings in research: Comment on friston (2012)". *NeuroImage,* ***81***(0), pp. 496 – 498.

[15] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R., 2013. "Power failure: why small sample size undermines the reliability of neuroscience". *Nature Reviews Neuroscience,* ***14***(5), Apr., pp. 365–376.

[16] Hubbard, R., and Bayarri, M. J., 2003. "Confusion over measures of evidence (p's) versus errors ('s) in classical statistical testing". *The American Statistician,* ***57***(3), Aug., pp. 171–178.

[2] `https://osf.io/hadz3/`

[17] Gigerenzer, G., 2004. "Mindless statistics". *The Journal of Socio-Economics,* ***33***(5), pp. 587–606.

[18] Freedman, D. A., et al., 2008. "Randomization does not justify logistic regression". *Statistical Science,* ***23***(2), pp. 237–249.

[19] Benjamini, Y., and Hochberg, Y., 1995. "Controlling the false discovery rate: A practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological),* ***57***(1), pp. 289–300.

[20] Austin, P. C., Mamdani, M. M., Juurlink, D. N., and Hux, J. E., 2006. "Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health". *Journal of Clinical Epidemiology,* ***59***(9), pp. 964 – 969.

[21] John, L. K., Loewenstein, G., and Prelec, D., 2012. "Measuring the prevalence of questionable research practices with incentives for truth telling". *Psychological Science,* ***23***(5), May, pp. 524–532.

[22] Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., and Wicherts, J. M., 2013. "Recommendations for increasing replicability in psychology". *European Journal of Personality,* ***27***(2), Mar., pp. 108–119.

[23] Fidler, F., 2002. "The fifth edition of the apa publication manual: Why its statistics recommendations are so controversial". *Educational and Psychological Measurement,* ***62***(5), Oct., pp. 749–770.

[24] Bakker, M., van Dijk, A., and Wicherts, J. M., 2012. "The rules of the game called psychological science". *Perspectives on Psychological Science,* ***7***(6), Nov., pp. 543–554.

[25] Freedman, D. A., 2010. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press.

[26] , 2013. "Announcement: Reducing our irreproducibility". *Nature,* ***496***(7446), Apr., p. 398.

[27] Ince, D. C., Hatton, L., and Graham-Cumming, J., 2012. "The case for open computer programs". *Nature,* ***482***(7386), Feb., pp. 485–488.

[28] Nature. Statistical checklist - Nature. `www.nature.com/nature/authors/gta/Statistical_checklist.doc`. Accessed: 2015-01-08.

[29] Baker, D., Lidster, K., Sottomayor, A., and Amor, S., 2012. "Reproducibility: Research-reporting standards fall short". *Nature,* ***492***(7427), Dec., p. 41.

[30] Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H. E., Golub, R. M., Goudreau, J. L., Gross, R. A., Gubitz, A. K., Hesterlee, S. E., Howells, D. W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S. E., Levine, M. S., Macleod, M. R., McCall, J. M., Moxley, R. T., Narasimhan, K., Noble, L. J., Perrin, S., Porter, J. D., Steward, O., Unger, E., Utz, U., and Silberberg, S. D., 2012. "A call for transparent reporting to optimize the predictive value of preclinical research". *Nature,* ***490***(7419), Oct., pp. 187–191.

[31] Humphreys, M., de la Sierra, R. S., and van der Windt, P., 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration". *Political Analysis,* ***21***(1), Jan., pp. 1–20.

[32] Open Science Collaboration, 2012. "An open, Large-Scale, collaborative effort to estimate the reproducibility of psychological science". *Perspectives on Psychological Science,* ***7***(6), Nov., pp. 657–660.

[33] Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A., 2012. "An agenda for purely confirmatory research". *Perspectives on Psychological Science,* ***7***(6), Nov., pp. 632–638.

[34] Stroebe, W., and Strack, F., 2014. "The alleged crisis and the illusion of exact replication". *Perspectives on Psychological Science,* ***9***(1), Jan., pp. 59–71.

[35] Johnson, D. H., 1999. "The insignificance of statistical significance testing". *The Journal of Wildlife Management,* ***63***(3), pp. pp. 763–772.

[36] Hoover, K. D., and Siegler, M. V., 2008. "Sound and fury: Mccloskey and significance testing in economics". *Journal of Economic Methodology,* ***15***(1), pp. 1–37.

[37] Orlitzky, M., 2011. "How can significance tests be deinstitutionalized?". *Organizational Research Methods*, p. 1094428111428356.

[38] Rozeboom, W. W., 1997. "Good science is abductive, not hypothetico-deductive". In *What if there were no significance tests?*, L. L. Harlow, S. A. Mulaik, and J. H. Steiger, eds. Lawrence Erlbaum Associates, New York, pp. 335–392.

[39] Haig, B. D., 2005. "An abductive theory of scientific method.". *Psychological methods,* ***10***(4), p. 371.

[40] Cash, P., Elias, E., Dekoninck, E., and Culley, S., 2012. "Methodological insights from a rigorous small scale design experiment". *Design Studies,* ***33***(2), Mar., pp. 208–235.

[41] Hernandez, N. V., Shah, J. J., and Smith, S. M., 2010. "Understanding design ideation mechanisms through multilevel aligned empirical studies". *Design Studies,* ***31***(4), pp. 382 – 410.

[42] Cash, P. J., Hicks, B. J., and Culley, S. J., 2013. "A comparison of designer activity using core design situations in the laboratory and practice". *Design Studies,* ***34***(5), pp. 575 – 611.

[43] Frey, D. D., and Dym, C. L., 2006. "Validation of design methods: lessons from medicine". *Research in Engineering Design,* ***17***(1), pp. 45–57.

[44] Kaptein, M., and Robertson, J., 2012. "Rethinking statistical analysis methods for CHI". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, ACM, pp. 1105–1114.

[45] APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008. "Reporting standards for research in psychology: Why do we need them? what might they be?". *American Psychologist,* ***63***(9), Dec., pp. 839–851.

[46] One, P. PLOS editorial and publishing policies. `http://www.plosone.org/static/policies#sharing`. Accessed: 2015-01-08.

[47] Freedman, D. A., 2009. "Diagnostics cannot have much power against general alternatives". *International Journal of Forecasting,* ***25***(4), pp. 833–839.

[48] Anscombe, F. J., 1973. "Graphs in statistical analysis". *The American Statistician,* ***27***(1), pp. pp. 17–21.

[49] Freedman, D. A., 2006. "Statistical models for causation". *Evaluation Review,* ***30***(6), Dec., pp. 691–713.

[50] Efron, B., 2010. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1. Cambridge University Press.

[51] of Health, N. I. NIH data sharing policy. `http://grants.nih.gov/grants/policy/data_sharing/`. Accessed: 2015-01-08.

[52] Vines, T. H., 2014. "Scientific community: Journals must boost data sharing". *Nature,* ***508***(7494), Apr., p. 44.

[53] Wilcox, R. R., 2012. *Introduction to robust estimation and hypothesis testing*. Academic Press.

[54] A, T., NG, B., and JA, I., 2007. "Persistence of contradicted claims in the literature". *JAMA,* ***298***(21), pp. 2517–2526.

[55] Freedman, D. A., 2009. *On Types of Scientific Inquiry: The Role of Qualitative Reasoning*. Cambridge University Press, Cambridge, ch. 20, pp. 337–356.

[56] Lab, A. D. A. Repository of design protocol data. `http://asudesign.asu.edu/protocol_repository/`. Accessed: 2015-04-09.

[57] Lab, D. E. Design repository. `http://design.engr.oregonstate.edu/repo`. Accessed: 2015-04-09.

[58] Fuge, M., and Agogino, A., 2015. "Pattern analysis of ideos human-centered design methods in developing regions". *Journal of Mechanical Design,* ***138***(4), July.

[59] Fuge, M., Tee, K., Agogino, A., and Maton, N., 2014. "Analysis of collaborative design networks: A case study of openideo". *Journal of Computing and Information Science in Engineering,* ***14***(2), Mar., pp. 021009+.

[60] Fuge, M., Stroud, J., and Agogino, A., 2013. "Automatically inferring metrics for design creativity". In ASME International Design Engineering Technical Conferences/DTM.

[61] Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., and Van der Laan, M., 2014. "Promoting transparency in social science research". *Science,* ***343***(6166), pp. 30–31.