Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency

Jan B. Vornhagen
jan.vornhagen@aalto.fi
Aalto University
Espoo, Finland

April Tyack
april.tyack@aalto.fi
Aalto University
Espoo, Finland

Elisa D. Mekler
elisa.mekler@aalto.fi
Aalto University
Espoo, Finland

ABSTRACT

Statistical Significance Testing – or Null Hypothesis Significance Testing (NHST) – is common to quantitative CHI PLAY research. Drawing from recent work in HCI and psychology promoting transparent statistics and the reduction of questionable research practices, we systematically review the reporting quality of 119 CHI PLAY papers using NHST (data and analysis plan at OSF.io). We find that over half of these papers employ NHST without specific statistical hypotheses or research questions, which may risk the proliferation of false positive findings. Moreover, we observe inconsistencies in the reporting of sample sizes and statistical tests. These issues reflect fundamental incompatibilities between NHST and the frequently exploratory work common to CHI PLAY. We discuss the complementary roles of exploratory and confirmatory research, and provide a template for more transparent research and reporting practices.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Transparency; Reproducibility; Open Science; Statistics

ACM Reference Format:
Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20), November 2–4, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3410404.3414229

Note: This systematic literature review was pre-registered at https://osf.io/z5ed2, following the PRISMA-P checklist [151].

1 INTRODUCTION

A primary goal of CHI PLAY is to provide a space for "*high quality research in games and HCI*" while "*embracing a wide variety of research contributions*" [3]. Many of these contributions emerge from empirical user studies of videogames and other game-like artefacts, whereby statistical analysis is applied to quantitative (or quantified) data to produce new insights regarding player-computer interaction [179]. Often, data analysis proceeds by way of p values (e.g., as computed via t-test or ANOVA), which are used to understand whether trends in data represent real effects, or merely noise. This is commonly called *Null Hypothesis Significance Testing* (NHST).

However, NHST methods have become increasingly subject to critique. False positive results, whereby noise is misidentified as a real effect, can easily occur as a result of common practices performed during analysis [79, 154]. These *Questionable Research Practices* [177, QRPs] threaten the legitimacy of statistical significance and therefore complicate interpretation of published research findings [79, 154]. QRPs are facilitated by a publishing climate biased towards statistically significant results[1], leaving non-significant research findings in the file drawer [33, 49, 131, 170].

A growing number of HCI scholars have consequently called for greater consideration of the quality of NHST analyses, and statistical reporting more broadly [26, 27, 48, 75, 88]. However, the extent to which these issues affect HCI research on games and play – and CHI PLAY in particular – is yet to be determined.

Yet CHI PLAY arguably has much to gain from other fields where similar problems have begun to be addressed. Quantitative games research in HCI often draws from psychological theory and methodology [e.g., 165] – as such, we argue that recent psychological work on Open Science has much to offer to the CHI PLAY community.

In this paper, we examine the quality of statistical reporting at CHI PLAY, reviewing 119 publications employing NHST in their analysis. We observe wide variation in reporting quality: about two thirds (67.2%) of these papers consistently report full test statistics, 28.6% contain inconsistent p values, and only seven papers (5.9%) justify their sample size. Moreover, NHST is often (mis)applied to exploratory research questions, risking the propagation of false positive findings. Our results demonstrate that quantitative research at CHI PLAY exhibits similar issues as other HCI research domains [26], suggesting a need to improve research practice, peer review, and publication guidelines. To help address these issues, we offer a comprehensive, easy to use template for authors and reviewers to assess the reporting quality of papers that employ NHST in their analysis. Moreover, we argue for the widespread adoption of Open Science practices, such as data sharing and pre-registration, by the CHI PLAY community.

In the following, we first describe the NHST method, and summarize pertinent issues with the approach identified by scholars in HCI [e.g., 26, 33, 48] and other fields [e.g., 79, 154, 158, 177]. We then present a systematic literature review of 119 papers employing NHST to evaluate the quality of statistical reporting at the venue. Lastly, we present a template for authors and reviewers, and

[1] A phenomenon we have also occasionally observed during peer review at CHI PLAY.

CHI PLAY '20, November 2–4, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-8074-4/20/11.
https://doi.org/10.1145/3410404.3414229

propose recommendations for future quantitative games and play research at CHI PLAY.

2 RELATED WORK

We begin with a brief summary of NHST methods, from which we situate our critique. A comprehensive overview is beyond the scope of this paper; however, for a more in-depth review of the underlying mathematical principles, we refer readers to [45, 48, 102].

Quantitative research usually takes one of two forms: *Exploratory* work, through which hypotheses are generated, or *confirmatory* studies, where hypotheses can be tested [48, 75]. Exploratory research is typically conducted in the early stages of forming theories and conceptual frameworks, where researchers are unsure what effects to expect. Data are collected and explored for potential effects, informing theory and concept development, which facilitate the generation of concrete hypotheses. The goal of confirmatory research, then, is to test these hypotheses through their use in predicting one or more study outcomes. In this way, correct predictions can represent support for a theory.

Researchers often test hypotheses using Null Hypothesis Significance Testing (NHST). NHST is an inferential approach to statistics that allows researchers to decide between two competing hypotheses about data: The *null hypothesis* (H_0) and the alternative hypothesis (H_1).

Researchers are usually interested in an entire population (e.g., online game players); however, it is usually impossible to test every person in a population of interest. Methods such as NHST can produce valid inferences regarding the complete population from a smaller sample.

However, extrapolating from incomplete data inherently comes with the risk of making one of two errors: falsely rejecting the null hypothesis when it is actually true (Type I), and failing to reject the null hypothesis when it is actually false (Type II) [88].

It is not possible to find support for a specific hypothesis with NHST, as p values represent the probability of observing at least equally extreme data under the null hypothesis. NHST instead allows for precise control over how often a hypothesis is falsely accepted or rejected *in the long run*. The significance (or α) threshold (usually .05) controls how often we will make a Type I error and incorrectly reject the null hypothesis (false positive). Statistical power ($1 - \beta$, usually set to .8) controls how often we will make a Type II error and incorrectly fail to reject the null hypothesis (false negative).

Using NHST, it is not possible to know whether a null hypothesis is truly false following a single significant result. Instead, NHST demonstrates the frequency of incorrect judgements in the long run when rejecting the null hypothesis. Because error rates are controlled over a series of hypothetically infinite studies, rather than just one, NHST is also referred to as a frequentist approach to statistics [45].

2.1 Criticism of NHST

While widespread, the use of NHST in HCI is controversial [33, 48, 88]. Two main issues emerge: First, a single p value constitutes only weak evidence; second, long-term error control can occur only when researchers abide by a number of specific rules. Error rates become inflated when these rules are not followed, which limits the validity of results. We discuss each issue in turn.

It is a common misunderstanding that p values represent the probability that a hypothesis is true [45, 48, 65, 88, 158]. More worryingly, p values are not a measure of strength of evidence like a standardized effect size, as they randomly fluctuate [88, 158] in what was coined the "dance of the p values" [34, 49]. For example, when performing t-tests on two groups with the same mean and standard deviation – in other words, when no true effect exists – all p values are equally likely; the proportion of significant p values is equal to the α threshold (i.e., 5%). Even when a true between-group difference exists, repeated tests will not produce the same p value: The p values "dance" around the significance threshold. This is illustrated in Figure 1.

This issue is compounded for studies with low *statistical* power. In Figure 1 left, power $\approx$ 80%; i.e., approximately 80% of p values correctly fall below the significance threshold, and "only" 20% of studies would commit a Type II error. The statistical power of a test is a function of both sample and effect size: power is maximized when a large effect is expected from a large sample. For underpowered studies, Type II errors become more likely, and null results consequently provide limited information[2] [88].

This behavior of the p value when the null is true – the uniform distribution in which every p value is equally likely – is problematic for exploratory study designs. More specifically, when more values are drawn from the uniform p distribution (i.e., as more statistical tests are run), the odds of observing spurious significant results are increased.

Exploratory research takes an essential role in HCI [33], as exploratory works where novel artefacts are designed and evaluated – often via a variety of measures and statistical tests – constitute key contributions to the field [179]. Yet applying NHST in exploratory research presents one of two drawbacks: if the critical α threshold is lowered to prevent an excess of false positive findings, potentially interesting (and non-significant) findings may be overlooked; conversely, not adjusting α substantially increases the Type I error rate such that confidence in test results is attenuated [35, 56]. Accordingly, exploratory work should be explicitly labelled as such, and indeed may benefit from using alternative analytic methods, such as estimation [48].

2.2 Questionable Research Practices and Researcher Degrees of Freedom

Sometimes researchers engage in exploratory analyses, yet report the resulting significant findings as if they were confirmatory – effects are claimed to support a hypothesis formulated after data collection has occurred. This practice is called hypothesizing after results are known, or HARKing [33, 91] and is a so-called *Questionable Research Practice* (QRP) [83]

QRPs are decisions made during data collection or analysis whose strategic application can improve the odds of achieving statistically significant results (and thereby inflate the Type I error rate). Typical QRPs include removing data points to change a group mean,

[2]It should be noted that we discuss *statistical* power, which is calculated with an expected effect size. So-called *post hoc* power, derived after analysis from the observed effect size and achieved sample size is merely a conversion of the p value and therefore of limited use and misleading [101].

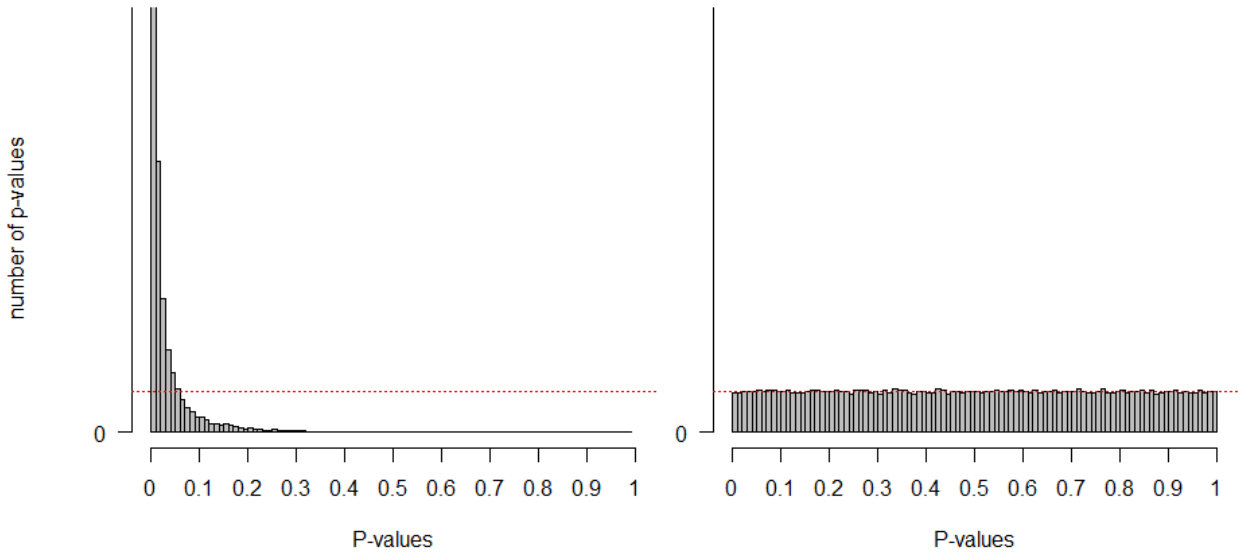

Figure 1: Histograms of p values, each produced from 100,000 simulated t-tests. For each t-test (n=50), samples were randomly drawn from two normal distributions. On the left, the two distributions had $m_1 = 100$ and $m_2 = 105$ ($sd_{1,2} = 10$); on the right, the two distributions were completely equal ($m_{1,2} = 100; sd_{1,2} = 10$). In both histograms, the first five bars include all p values < 0.05, constituting significant results. Note that ~5% of results in the right distribution were significant. The red bar represents the uniform distribution of p values for infinite simulations where the true effect is zero. Plots generated with code by [102] (CC BY-NC-SA 4.0).

collecting data until desired results occur, selectively reporting (in)dependent variables, rounding down p values close to the significance threshold, and HARKing [83, 142, 177].

These practices may, at first glance, seem reasonable: for example, continuing data collection when analysis yields non-significant results [83] does increase the sample size, which is generally desirable. However, because p values behave randomly (Figure 1), these further tests always increase the chance of false positive findings. In fact, this practice of intermediate testing until a significant result is found will eventually yield significant results regardless of what is tested [158].

It has been shown that QRPs are pervasive in the psychological literature [83]. Their impact can be immense: Only reporting one of two dependent variables, collecting ten more observations per cell, controlling for gender, and selectively choosing between three conditions collectively increases the rate of false positive findings from 5% to 60% [154][3]. The influence of QRPs has been deemed so massive that in a widely discussed paper, Ioannidis [79] proposed that most published research findings are false, with some scientific fields potentially only reporting their own biases instead of any true effects.

A number of approaches to prevent QRPs have been identified [33]. One prominent option is pre-registration, whereby scholars record all salient features of their research plan in a time-stamped, immutable form prior to data collection [33, 122]. While the exact contents of pre-registrations tend to differ [see 122], they typically include the research goals, hypotheses to be tested, and the statistical analysis plan [33]. Pre-registration plans represent compelling evidence that confirmatory statistical analyses are independent of the observed data. As such, pre-registration practices represent a positive step towards more open and transparent science; however, they do not constitute an absolute safeguard against QRPs. Pre-registered studies lose value when research plans omit key information, or differ from final analyses without a valid rationale [177].

QRPs are made possible by the substantial amount of leeway afforded during research, otherwise known as Researcher Degrees of Freedom (R-DFs) [154, 177]. These freedoms are not bad in and of themselves: many decisions, such as which participants to exclude from a sample, are essential aspects of research (for example, to ensure data integrity [23]). However, because researchers are incentivized to produce "novel" (i.e., significant) results for publication [33], R-DFs can be readily misused to "improve" results [83]. Conversely, many non-significant results are never published for the same reason, resulting in the *file drawer problem* [33, 137].

We note that QRPs do not constitute fraud. Practices such as intermediate testing have historically been considered unproblematic and defensible [83]. In contrast, fraud requires intent, and is

[3] An interactive example can be found in the P-Hacker app, which illustrates this process [144]

therefore rare. QRPs are widespread, easily performed unintentionally, and difficult to detect, making them a primary threat to the reproducibility of results [142, 154, 177].

R-DFs serve a valid purpose in the scientific process – for example, excluding careless survey respondents can improve data quality [23]. However, their application should be intentional, consistent, and transparently reported. To these ends, Wicherts et al. [177] compiled a list of 34 R-DFs to help researchers identify unintentional *p*-hacking in their own practices; for example, testing broadly stated hypotheses for which a number of dependent variables *could* apply – or, if non-significant, be removed from the analysis.

In the present work, we apply this list of 34 R-DFs to evaluate the quality of statistical reporting at CHI PLAY. In doing so, we aimed to understand how R-DFs at CHI PLAY are reported, and hence to what extent the rate of false positive results has been inflated. In this way, the present research follows from a long tradition of meta-scientific work on research methods at CHI PLAY, CHI, and the wider HCI literature.

Calls for other changes to research methods and reporting practice have also been made in recent games scholarship. The use of non-violent "control" videogames in aggression research [8], for example, provoked questions as to whether games that vary across genre, pacing, and content can induce comparative experiences [51, 73]. More detailed reporting of game selection procedures, with theoretical or empirical bases, has been suggested as a means by which researchers could more convincingly justify their choice of stimulus games [166]. Psychometric measures have also come under increasing scrutiny in HCI games research. In particular, it was shown that the Game Experience Questionnaire (GEQ) saw wide use as a validated measure, despite the absence of a published validation study [109] – and indeed, the stated factor structure could not be independently validated [24, 85, 109]. Moreover, substantial variation in reporting basic qualities of the GEQ, such as the number of scale items, was observed across the literature [109], suggesting the existence of broader issues in reporting practice in HCI games research.

3 SYSTEMATIC LITERATURE REVIEW

We conducted a systematic literature review to take stock of the quality of statistical reporting at CHI PLAY. Following the PRISMA-P protocol [151], the literature review was pre-registered on January 29 2020, before beginning data collection. Materials and the PRISMA Flow diagram detailing all steps of the review are available at https://osf.io/4mcbn/. The literature review and analysis were performed by the first author, in regular consultation with the third author.

3.0.1 Identification. Using the ACM Digital Library[4], we collected all CHI PLAY papers published since the inaugural conference in 2014 that were classified as "Research-Article" (i.e., full papers). As such, we did not include publications labeled as "Abstract" or "Short-Paper", as the reduced page limit and potential preliminary status of the work (e.g., Works-in-Progress) may have limited what authors could report. This first step resulted in a sample of n=246 papers.

[4] dl.acm.org/conference/chi-play/proceedings

3.0.2 Screening. Next, we screened the sample of 246 papers for Notes. We decided to exclude Notes from our sample, for the same reason that we excluded Abstracts and Short-Papers. We excluded n=8 notes, resulting in n=238 considered for further analysis.

3.0.3 Eligibility. We screened the remaining papers for the presence of inferential statistics; for example, reporting *p* values or describing results as significant. In this way, a further n=108 papers without inferential statistics were excluded, leaving a sample of n=130 papers (marked with * in the References).

3.0.4 Codebook. Initial coding proceeded by adapting the checklist developed by Wicherts et al. [177] into a preliminary codebook. These early categories included R-DFs organized around Hypothesizing (e.g., "Conducting explorative [sic] research without any hypothesis"), Study Design (e.g., "Measuring additional constructs that could potentially act as primary outcomes"), Data Collection (e.g., "Determining the data collection stopping rule on the basis of desired results or intermediate significance testing"), Statistical Analysis (e.g., "Choosing to include different measured variables as covariates, independent variables, mediators, or moderators"), and Reporting of Results (e.g., "Failing to assure reproducibility (verifying the data collection and data analysis)") [all quotes 177, Table 1, p. 3].

As per our pre-registration report, we randomly selected a subset of papers (n=32; ~25% of the sample) to assess the preliminary codebook's viability. Based on this analysis, the codebook was revised in several ways:

- As only two studies in our sample were pre-registered (i.e., [81, 175]), R-DFs that could only be identified with knowledge of authors' pre-study intentions were removed (e.g., post-hoc switching of the primary outcome).
- As we will elaborate in our Results section, hypotheses, test statistics, effect sizes, measures, and assumption tests were often reported incorrectly or not at all (e.g., hypotheses were only implicitly linked to tests performed). We therefore added codes referring to complete and clear reporting practices (e.g., "are the (in)dependent variables reported in a way that readers could reproduce them?").
- We added items for statistical reporting (e.g., "are full test statistics reported?") to address concerns previously noted in Related Work regarding the improper use of NHST. We also examined *p* value reporting practices with statcheck.io [121], or manual computation where necessary. In particular, we investigated whether reported *p* values were consistent with their corresponding degrees of freedom and test statistics.
- We added items related to reproducibility (e.g., "does the paper provide raw data or an analysis plan?") to assess to what extent Open Science practices, such as sharing data, have been adopted at CHI PLAY.

The final codebook, which includes code descriptions and relevant citations, can be found in the OSF repository.

3.0.5 Included. We analyzed all 130 papers with the updated codebook. During this final analysis, a further n=11 papers were excluded: Four papers did not feature statistical inferences, and were therefore deemed false positives [15, 31, 47, 74]; two studies used exploratory factor analyses [109, 161], which (unlike NHST) are

designed for exploratory analyses; finally, five papers did not employ NHST when reporting results [125, 171, 173], or only reported non-significant results without further information on the statistical analysis [119, 175]. This resulted in a final sample of N=119 papers, which forms the basis of our systematic review results.

4 RESULTS

In the following section, we present the results of our literature review. Altogether, our sample spans almost half (48.78%) of all CHI PLAY full papers published between 2014 and 2019, attesting to the popularity of NHST at the conference. We present findings in the general order of our codebook. Importantly, our review concentrates on the quality of the (described) *methods and reporting*; we do not intend to make statements on the research quality of the works.

4.1 Hypothesis Reporting

As noted in Related Work, long term error rates are only controlled in confirmatory research designs that test specific statistical hypotheses. We therefore examined whether papers employing NHST reported confirmatory research goals and statistical hypotheses.

Most works contained at least one exploratory research question (n=75, 63.02%), though it was not always labeled as such. For some studies, however, an exploratory focus was explicitly noted; for instance, "in the absence of an existing theoretical framework for the design and discussion of asymmetric games, we adopted an exploratory approach" [70, p. 350]. Other papers were less clear: in some cases, no research questions were identified [e.g., 13]; in others, research goals were described in ways that could not be interpreted as confirmatory [e.g., 11] – for example, among one study's "primary aims" [14, p. 327] was to "[d]etermine the similarities and differences of the likely impacts of MDDA [...]" (p. 327). Note that while this statement is a useful research question and may be considered a "theoretical" hypothesis, it does not constitute a "statistical" hypothesis or confirmatory research goal, as it does not directly relate to test outcomes.

Of the 119 papers, n=44 (36.98%) stated a confirmatory research goal, outlining at least one hypothesis that was supported or rejected on the basis of their results. For example, two hypotheses are defined in [41], one of which predicts that "[e]xperiences of interdependence (H1a) and cooperation (H1b) are positively associated with in-game social capital" (p. 90). As both H1a and H1b are later defined in terms of self-report measures, this hypothesis can be directly tested.

Notably, over half of the reviewed works (n=64, 53.78%) stated hypotheses (e.g., "There will be no differences in any game experience measures due to graphical fidelity", [21, p. 270]), even where confirmatory research goals had not been formulated. Only n=22 (18.49%) outlined statistical hypotheses that could be directly translated into statistical tests (e.g., "The perspective switching provides significant benefits to spatial orientation and overview" [32, p. 290]). This is mirrored in the reported analyses, where for most papers (n=90, 75.63%), the distinction between confirmatory and exploratory testing is unclear. In comparison, 12.61% (n=15) of papers clarified this distinction – for instance, by using subheadings (e.g., "Exploratory Analysis: Game Experience and Game Behavior", [124, p. 9]; "Further Statistics", [185, p. 7]). In the remaining 11.77% papers (n=14), only confirmatory analyses were conducted.

We were unable to confidently assess whether tests for all hypotheses were reported in 50.42% (n=60) of papers, either because the hypotheses themselves were not clearly stated [e.g., 37], or the reporting did not allow us to assess whether a hypothesis had been answered: For example, one paper sought to study how a system might "collect en-masse achievement data about gamers" [p. 305 174], "[w]hat insights [might] be drawn solely from the data collected [...]" (p. 305), and identify the limitations of "using only one source of usernames for the system [...]" (p. 305). While the first two questions were addressed in the results section, the open wording of the hypothesis makes it impossible to determine whether the questions were addressed completely. Indeed, the third research question is not discussed in the results section at all, but only addressed in the limitations.

In the remaining 49.58% (n=59) of cases, all hypotheses could be clearly linked to a corresponding test [e.g., 72, 157].

4.2 Study Design Reporting

Researchers have the most freedom when planning and preparing a study. Detailed and thorough reporting of the study design is therefore crucial for readers to understand how a study was conducted, the ways its enactment may have influenced results, and how it could be replicated.

4.2.1 Sample Size. An important part of NHST is justifying the sample size, usually via power analysis and a defined significance threshold (e.g., $\alpha < .05$). Of the N=119 papers, n=7 (5.88%) justified their sample size [e.g., with a power analysis, as in 44], with one of the seven only reporting post-hoc power [21] – which is somewhat misleading, as post-hoc power is not equivalent to statistical power, but rather a conversion of the p-Value [101].

All papers reported the sample size, where n=54 (45.38%) either did not remove any participants from their analysis [e.g., 9], or provided a rationale for doing so [e.g., where data collection was compromised for participants who had guessed the true intent of the study, as in 168]. A further 21.01% (n=25) of papers did not justify removing participants, or did not mention having done so, despite inconsistencies between reported degrees of freedom and sample size [i.e., indicating that participants were removed from the test, e.g., 6, 28]. In the remaining 33.61% (n=40) papers, insufficient details were reported to determine whether participants were removed.

4.2.2 Significance Threshold. The significance threshold (i.e., α) was defined more frequently, with 18.49% (n=22) of papers opting for a single value (i.e., 0.05) for all tests. All other papers either used multiple thresholds or an implicit threshold. Further, only 4 papers (3.36%) justified their α threshold. For example, [164] adjusted the standard α threshold from 0.05 to 0.0045 "[...] to control the experiment-wise error rate across the 11 tests [...]" (p. 5).

4.2.3 Study Setup and Variables. A clear description of the study design and setup are needed to understand how the study was conducted, what was measured and how, as well as how the researchers managed their degrees of freedom. Independent variables were thoroughly described in most papers (n=92, 77.31%), resulting

in a clear vision of the manipulation, and facilitating conceptual replications.

Dependent variables were defined in less detail. Overall, 56.3% (n=67) papers reported them in ways that clarified their role in the study and made them available for replication. For example, one paper used a subheading per questionnaire, stating "To quantify the play experience, we measure interest/enjoyment, invested effort and pressure/tension using the IMI" [42, p. 453], followed by a citation.

In papers where dependent variables were not precisely defined, the methods of their construction, relevance to the research, or potential moderator status was rarely stated [e.g., 54]. A majority of studies reported additional measures: in 51.26% (n=61) of papers, these variables were not described in a way they could be replicated or their role in the analysis was unclear (e.g., whether they were intended as dependent or moderator variables). Of the remaining papers, 14.29% (n=17) fully reported their additional variables, and 34.45% (n=41) did not report measuring additional variables. Most papers (n=88, 73.95%) did not employ moderator variables, but among those that did, few explained their use (n=14, 11.77%).

An exemplar of reporting measurements and manipulation can be found in Johanson et al. [81]. Moderator variables were collected "to get a sense of each participant's interest in the task and ability to complete the task" [81, p.174], and are subsumed under the "Questionnaires" subheading. Dependent variables are separately described in the next subsection "Dependent Measures".

4.3 Statistical Reporting

Of the 119 papers, n=80 (67.23%) report their tests in a way that communicates (1) the tests used, (2) degrees of freedom, (3) the test statistic, and (4) the p value. In the remaining papers, at least one of these elements was not clearly reported. An example of a well-reported ANOVA that also incorporates effect size can be found in [114, p. 196]: *"(F(2, 122) = 56.8, p < .001, η_p^2 =.482)"*.

When sufficient statistical details were reported, we used the app statcheck.io [121] to review the computation of p values from the test statistic and degrees of freedom. As statcheck only works for papers that report results formatted according to APA guidelines (and PDFs that directly translate into plain text), results were computed by hand where necessary. While 51.26% (n=61) of papers reported consistent p values, inconsistencies were observed in 28.57% (n=34) of papers. In most cases, inconsistencies reflected rounding errors with no meaningful influence on study outcomes – rarely, however, we observed decision inconsistencies, whereby the reported and computed p values supported different decisions. For 20.17% (n=24) of papers, it was not possible to re-compute p values, as they lacked necessary statistical details.

For example, one paper reported "$t(14) = -2.055, p = 0.049$" [22, p. 211], which would result in a significant $p = 0.0295$ for a one-sided test, or a non-significant $p = 0.059$ if a two-sided test was conducted. The same paper reported "$F(2, 12) = 3.775, p = 0.031$" (p. 211), which would produce $p = 0.053$ [5]. Note that these were two tests among a total of 25 reported in the paper. As such, the observed inconsistencies do not change the overall conclusions of the paper by much.

4.3.1 Assumption Testing. Most papers (n=108, 90.76%) reported parametric tests, which should generally be accompanied by assumption tests – however, the majority of papers (n=64, 53.78%) did not mention these. Of the remaining papers, 36.98% (n=44) explicitly described assumption tests [e.g., 68, 69], the remaining 9.24% (n=11) used non-parametric tests [e.g., 160] or tests, where we were not aware of applicable assumption tests [e.g., 19].

4.3.2 Effect Sizes and Confidence Intervals. We also examined the prevalence of reporting effect sizes and confidence intervals (CIs): While we found effect sizes in 63.02% (n=75) of papers, only 6.72% (n=9) reported CIs [e.g., 58, 86] for point estimates of interest (usually effect sizes or means), suggesting that many studies rely solely on p values for their inferences. Lastly, few papers adjusted their significance threshold for multiple testing: 81.51% (n=97) of papers did not report adjusting their critical α level despite conducting multiple tests related to one hypothesis. Of the remaining works, 10.92% (n=13) performed adjustments [e.g., 114, 138], and a further 7.56% (n=9) did not require adjustment for multiple tests [e.g., 62, 94].

4.4 Transparency

Transparent data, analyses, and research goals allow other researchers to independently reproduce the analysis, or perform replication studies.

Only n=5 papers (4.2%) were accompanied by publicly available data [e.g., 40, 148, 160]; the remaining papers did not provide explanations for their non-disclosure. No papers shared the software script used for data analysis.

With regards to sharing experimental software, tools, or other materials, 20.2% (n=24) of papers sourced all relevant materials [e.g., 163], attaching questionnaires to the work [e.g., 162, 181], describing materials exhaustively [e.g., 78], or making software available in a repository [e.g., 145, 147, 148].

Finally, we investigated study pre-registration. Only one[6] paper [81] was pre-registered; curiously, the methods and tests described in the paper sometimes diverged from the pre-registration without explanation.

5 DISCUSSION

The present work has reviewed 119 CHI PLAY papers employing NHST to examine the quality of statistical reporting practice. We have identified a number of issues with the ways that study design and data analysis are reported in these papers. NHST is an extremely popular analytic method at CHI PLAY, with our corpus comprising almost half of all published full papers from the venue.

However, critiques of NHST have emphasized the ease by which false positive results can emerge from seemingly reasonable practices conducted during and after data collection [i.e., QRPs, 177]. Our review raises similar concerns about quantitative research at CHI PLAY.

[5] This was determined with statcheck.io, and rechecked via www.socscistatistics.com/pvalues/tdistribution.aspx and www.socscistatistics.com/pvalues/fdistribution.aspx. Both methods produced consistent p values that differed from those stated in the text.

[6] We note another pre-registered study [175] published at CHI PLAY 2019. However, the study was excluded from our review, as it describes a qualitative, exploratory approach and did not employ NHST.

Two main issues can be identified from our review. First, statistical reporting varies widely: unexplained changes to sample size between tests, uncorrected multiple testing, and vaguely specified dependent variables are common. Inconsistent reporting problematizes evaluation of research quality, potentially obscuring questionable practices (e.g., incomplete test statistics). This ambiguity can drastically increase the rate of false positive results in the literature.

Second, we identified a number of arguably exploratory studies that apply methods intended for use in confirmatory research. These papers often *resemble* confirmatory work, featuring hypotheses that are tested using NHST. However, these papers often reflect an exploratory intent; they may, for example, employ a wide array of measures to evaluate a game (or game-like artefact) over a similarly extensive battery of tests. Without correcting for multiple tests, this approach inflates the rate of false positive findings.

Although some works report their results to a high standard, in general, the perfunctory application of NHST in CHI PLAY research is a cause for concern. The pursuit of meaningful claims about player-computer interaction – "discussion of current high quality research in games and HCI" [3] – is impeded by a quantitative literature for which effect sizes and confidence intervals are commonly elided in favour of isolated p values, and whose analytic methods fundamentally conflict with high-level goals of research.

5.1 The Value of Confirmatory and Exploratory Research

The majority of reviewed papers contained no confirmatory research goal, and had not formulated any statistical hypotheses. This may suggest that (1) most of these works actually pursue exploratory research aims, even if not explicitly stated in the paper; and that (2) confirmatory research is seemingly of limited use in the context of player-computer interaction.

Indeed, "*intentionally* exploratory studies are a cornerstone of HCI" [33, p. 5, emphasis added]. As such, the prevalence of exploratory work at CHI PLAY is unsurprising. Many publications describe the testing phase of iterative development, and evaluate novel interfaces [33]; for these applications, the rigor of confirmatory approaches is not always needed [76]. Moreover, in contrast to experimental psychology or medicine, player-computer interaction is yet a nascent field of research, in which the discovery of novel phenomena for theory-building remains a priority. Phillips et al. [129], for instance, formulated an exploratory research question to investigate how their reward taxonomy affects the player experience (i.e., "RQ1: Does type of video game reward in a game influence the player experience?", p. 396). As such, exploratory research is well-suited to investigating topics that are not sufficiently understood, or where firm predictions are impractical.

What value, then, does confirmatory research have to CHI PLAY? Recall that confirmatory studies test hypotheses derived from theoretical and conceptual speculation – in other words, they *build on prior work*. For example, the qualitatively greater autonomy identified in solitary play, relative to social play with friends [169] could be formally tested from a confirmatory perspective. While confirmatory research is less common in HCI, its relative absence has contributed to concerns regarding fragmentation and limited progress [75, 88]. Hence, confirmatory research is necessary to advance player-computer interaction by linking empirical work to theoretical considerations, validating conceptual assumptions (e.g., greater variation in rewards increases intrinsic motivation [cf. 129]), and understanding the processes involved in particular phenomena [75]. Together, these efforts facilitate more informed predictions, as well as contribute to a more unified and integrative understanding of player-computer interaction.

5.2 Directions

Our review has identified a slew of shortcomings in current research and reporting practices at CHI PLAY. However, we emphasize that the QRPs described in the present work are likely a product of unfamiliarity with the assumptions underlying NHST, rather than intentional data massaging. As noted, many R-DFs have only recently been identified as problematic [83].

We highlight that some CHI PLAY research is, at times, already conducted in partial alignment with Open Science principles. Schwind et al. [148], for example, made their *faceMaker* app freely available, facilitating its further use in research. Similarly, despite concerns regarding its arguably low statistical power, Johanson and colleagues' [81] thorough pre-registration and detailed reporting clarify the aims of the work and facilitate replication.

Finally, we highlight that while essential, reporting quality is not the *sine qua non* of publication value – indeed, there are many other aspects (e.g., subject matter, theory) that make papers interesting and worth reading [75, 179].

However, for CHI PLAY research to have proceeded largely in isolation from these discussions is worth further examination. Guidelines for high-quality study design [154, 177], analysis [158], and reporting practices [4] – many of which were compiled from within HCI [27, 33, 48, 75, 88, 170] – have been largely eschewed. While we can only speculate as to why these recommendations have not yet found their way to the field, we urge CHI PLAY scholars to engage with works such as these, and more completely apply these guidelines in their own practice. Moreover, we recommend that HCI and games educators sensitize students to the different roles of confirmatory and exploratory research.

6 A TEMPLATE FOR MORE TRANSPARENT QUANTITATIVE RESEARCH AT CHI PLAY

To facilitate more rigorous and transparent research and reporting practices at CHI PLAY, we contribute a template for researchers to guide their study designs from start to finish, and for reviewers to quickly assess reporting quality. Recommendations mostly focus on confirmatory, NHST-based research, but also include pointers for exploratory work. Suggestions are presented alongside their corresponding citations for further reference.

Note that this template is neither complete nor infallible – research is diverse, and no single template can perfectly address every work – however, it is intended to specifically address concerns of quantitative studies published at CHI PLAY at present. As with p values and effect sizes, researchers and authors should take a critical perspective and carefully consider the rationale for each point to determine where divergence is pertinent.

6.1 Deciding on the Research Goals

First, researchers need to consider their research question(s) of interest, and how these may be studied [75]. Specifically, whether their aim is to test specific hypotheses (confirmatory), collect rich descriptions of a phenomenon or artefact (exploratory), or a combination thereof (i.e., confirmatory hypothesis testing, followed by exploratory analyses). This decision then informs all subsequent methodological choices, as well as the selection of suitable statistical methods.

6.2 Hypothesizing

Designing a study should proceed with the research goals in mind, as they influence a number of decisions that follow. Submissions should clearly report the overarching research goal, research questions, and precise hypotheses.

- **Confirmatory Research Goal [4, 177]:** The work has a clearly stated confirmatory research goal: Authors specify two competing hypotheses. Deciding on a clear research goal early on can drastically improve the study design, as it informs all subsequent study design choices.
- **Precise & Directional Statistical Hypotheses [88, 177]:** Hypotheses should be explicitly defined, including predictions as to how the independent variable will impact specific measures. Steinemann et al. [157], for example, built on previous experiments in media psychology to formulate concrete statistical hypothesis, e.g., "H1: Interactivity will lead to increased donations" [157, p. 321].

6.3 Study Design

Designing the study follows directly from the research question(s).

- ***Independent* variables are precisely defined [4, 154, 177]:** Independent variables should be clearly described and justified to facilitate replication. For example, when selecting video games as stimuli for experimental conditions, clarifications should be provided to justify the choice [166].
- ***Dependent* variables are precisely defined [177]:** Full reporting of dependent variables paints a more complete picture of the research (and results), and facilitates replication. Authors should describe which variables they intended to influence, how these were measured, and why this instrument was chosen. Especially in light of the variety of available player experience questionnaires, psychometrics and a clear rationale should be provided [85, 109].
- **All additional and moderator variables are defined [177]:** *All* additional variables collected are clearly described, including their role in the analysis. Demographic details and other interesting constructs with no clearly specified relation to the research question may provide valuable insights, but unless specified in the hypotheses, should not be used in confirmatory analyses.
- **Data cleaning, exclusions and grouping [177]:** All data cleaning practices, exclusion of participants' data, and grouping participants by demographic variables should be clearly summarized [see also 23], alongside a rationale for their use. Justifying these decisions (e.g., listing predetermined exclusion criteria) emphasizes that these measures were not taken post-hoc to fish for significant results. Where participants are grouped via a third variable (e.g., age brackets), a multiverse analysis – in which all reasonable groupings are calculated [49] – may help researchers demonstrate the robustness of their decisions.
- **Data collection and power [158, 177]:** Controlling for statistical power is crucial to understanding the long-term error rate. This is usually done by collecting a sample of a specific size, as determined by a power analysis. Power analysis also helps reduce over-testing (i.e., potentially wasting resources), and facilitates the development of more specific hypotheses, as it requires a prediction concerning the expected effect's magnitude (i.e., effect size).
- **Deciding on an alpha threshold [45, 104]:** As with statistical power, a single critical alpha value should be determined prior to data collection. The alpha threshold should be considered strictly dichotomous: Marginally significant results (typically $0.05 \leq p \leq 0.1$) should not be interpreted [130]. While $\alpha < 0.05$ is standard, the chosen threshold value may differ depending on the needs of the research [104].
- **Visualize the study design [50]:** Tools exist to walk through the aforementioned steps, visualize the study design, and facilitate pre-registration. Touchstone2 [50], for example, allows researchers to set up and compare study designs, and perform power calculations.

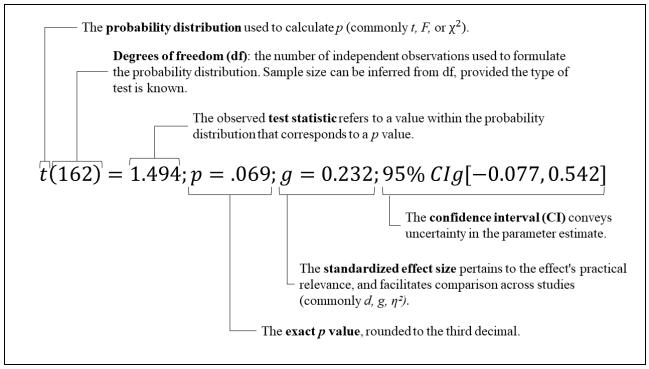

Figure 2: An illustration of a full test statistic. It includes all information the reader needs to understand which test was performed, how many data points were included, what the result is, and how the result may be interpreted.

6.4 Reporting

The aim of NHST is to decide between two competing hypotheses about data. Transparent reporting is necessary to comprehend these decisions, as well as promote reproducibility.

- **Full test statistics are reported [177]:** Statistical tests, degrees of freedom, and statistical values are reported in detail. A (fictitious) example of a fully reported one-sided, non-significant Student's t-test is depicted in Figure 2.
- **Assumptions are tested and reported [26, 177]:** All statistical tests come with specific assumptions towards the data. While some tests (e.g., a one-way ANOVA) are considered robust to violations of normality [143], other assumptions

such as variance homogeneity can affect common NHST tests, including ANOVA and ANCOVA [135], especially for smaller sample sizes. Papers should also clearly report if assumptions were not tested [e.g., tests of the normality assumption have been criticized for their unreliability 27].

- **Reporting effect sizes and confidence intervals [75, 90, 158]:** In contrast to p values, effect sizes indicate the strength of a statistical effect (e.g., how strongly the independent variable impacted the dependent variable), and provide a more meaningful basis for interpreting results. Confidence intervals should also be reported to indicate the degree of uncertainty around important point estimates (i.e., effect sizes, means, etc.). Reporting effect sizes and CIs also allows for comparisons across studies (e.g., as in meta analyses).
- **Correcting for multiple tests [16, 103]:** Family-wise error control is important to adjust for multiple testing, as every test beyond the first increases the rate of false positive findings. As a general rule, p values should be adjusted per hypothesis, where given k independent tests, the chance of observing a false positive is $1 - (1 - \alpha)^k$ [16]. Harpstead et al. [69], for example, chose a Bonferroni correction to adjust their chosen significance threshold of 0.001 "by the number of statistical tests being performed, (1 + the number of rewards) × (1 + the number of groups) × 2, resulting in a final alpha value of 5.05E^{-6}"[69, p.375].
- **Dealing with non-significant results [46, 130]:** As noted in Related Work, p values do not represent the probability that a hypothesis is true. Equally, non-significant results cannot be interpreted as evidence of "no effect": Any series of tests is likely to produce some non-significant results when statistical power is < 1. "Marginally significant" results are non-significant, and should not be interpreted otherwise [130].
- **Matching hypotheses to tests [177]:** All hypotheses should be clearly linked to corresponding statistical tests, as well as clearly report any exploratory and post-hoc analyses.

6.5 Transparency

Public sharing of study data and materials increases transparency and trust [116, 117], while freeing up space in papers to concentrate on other relevant aspects (e.g., the game design process) and discuss the most interesting findings.

- **The data set is available [4, 116, 117, 170, 177]:** If possible, anonymized raw data should be made openly available in a persistent online repository. The location of the data should be explicitly noted in the text, preferably in the abstract, so that the data are available, even if the paper is not openly accessible. Where anonymized data cannot be made public (e.g., vulnerable populations), researchers could instead generate a distributionally identical 'synthetic' data set [see 132, for a primer].
- **The analysis plan is available [4, 170, 177]:** Sharing R scripts, SPSS syntax, or the data analysis plan allows researchers and reviewers to easily replicate the analysis.
- **Experimental artefacts are available [4, 170, 177]:** All materials necessary to replicate the study should be made openly available. For many CHI PLAY publications, this simply entails listing all questionnaire items, as well as sourcing or uploading all study materials. Where this is not possible (e.g., due to copyrighted hardware, software prototypes), researchers should share alternate resources to facilitate replication. Krekhov et al. [98], for instance, describe a blueprint for their controller prototype.

6.6 Pre-registration

As all aforementioned recommendations refer to decisions made prior to data collection, pre-registering a study requires little extra effort [158]. The pre-registration plan provides readers a means to identify confirmatory and exploratory goals of the work, and discourages HARKing [33].

- **Pre-registration [4, 33, 117, 131, 170, 177]:** The study is pre-registered at a permanent third-party archive (e.g., OSF.io or AsPredicted.org) and accompanied by timestamps. Johanson et al. [81], for instance, provide an example of a pre-registration plan in the context of player-computer interaction.
- **Deviations from pre-registration are justified [33, 117, 131, 170, 177]:** Studies do not always go as intended. Deviations from the pre-registration plan are often warranted, but should be clearly highlighted and justified in the paper.

6.7 Exploratory Research

The previous sections reflect recommendations primarily tailored to confirmatory work and NHST. However, in light of the importance of exploratory research for player-computer interaction, we provide a few pointers here for more transparent quantitative exploratory research.

- **Stating the exploratory research goal [33, 177]:** Intentional exploratory analyses are perfectly reasonable, when clearly declared as such. Researchers should specify their exploratory focus, and avoid presenting findings as definitive proof (e.g., via p values). In turn, reviewers should not insist on the provision of p values, as they tend to be misconstrued as strong evidence.
- **Adequate statistical approach:** Some research questions may be more appropriately investigated with other statistical methods. When relevant prior information exists, Bayesian approaches [100] may be useful; in other cases, researchers may benefit from estimation-based approaches [48], as in some existing CHI PLAY work [e.g., 47].
- **Interpreting results [36, 48]:** Exploratory analysis is less constrained than a confirmatory approach. Hence, researchers might forgo certain "rules" when interpreting and reporting results. For instance, instead of relying on p values, researchers may gauge graphs "by eye" to compare whether confidence intervals overlap between groups.
- **Transparent reporting:** Exploratory work should also follow the transparency guidelines described above, including pre-registration [33], open data, and complete reporting of all collected measures. Sharing the data openly also allows other researchers to engage in exploratory analyses, and extends the contribution of the original work [170].

7 LIMITATIONS AND FUTURE WORK

First, it was not always obvious to determine whether the methodological best practices were followed in the reviewed papers. We could, for example, only investigate outlier removal based on the descriptions and rationales reported in the papers. Had the raw data been made openly available, it would have been possible to reproduce by what means outliers were identified and removed.

Second, screening and coding processes were conducted entirely by the first author. While these procedures were performed with transparency and completeness; however, some readers may disagree with aspects of the analysis.

Third, some points of critique around NHST remain topics of active debate; for example, whether and how to test assumptions [26, 27, 143], and when to adjust for multiple testing [103]. While we have made clear recommendations for many of these topics, we urge CHI PLAY researchers to more deeply engage with these discussions, to make informed decisions regarding their statistical analyses, and provide clear justifications in their papers.

Finally, our review is limited to NHST, researcher degrees of freedom, and their potential inflation of the Type I error rate. Yet our findings showcase further methodological concerns (e.g., suitability of statistical tests or study designs for answering research questions) that warrant consideration. Moreover, the present work mostly focuses on hypothesis-testing. However, recent work on the uses of psychological theory in HCI games research [165] also suggests a need to assess *how* hypotheses are generated and *what* research questions are formulated [75, 115].

8 CONCLUSION

Null Hypothesis Significance Testing (NHST) is a popular analytic tool at CHI PLAY. However, our review of 119 full papers highlights a number of inconsistencies and shortcomings with regards to research and reporting practices. These issues emerged against a backdrop of systematic misuse of confirmatory methods, such as NHST, in seemingly exploratory work. To help counter these issues, we present a template for authors to improve study design and statistical reporting, and for reviewers to evaluate work employing NHST. We are confident that by adopting basic Open Science standards – such as pre-registration, open data, and more uniform statistical reporting – the quality of CHI PLAY research may be further improved, fostering the validity and reliability of research findings.

9 DATA AVAILABILITY STATEMENT

The data, full analysis, as well as supplementary materials are available at https://osf.io/4mcbn/.

10 DECLARATION OF CONFLICTING INTERESTS

The authors declare no conflicting interests.

11 AUTHOR CONTRIBUTIONS

JBV and EDM conceptualized the paper and designed the study. JBV conducted data collection and analysis. EDM and AT consulted on the analysis. EDM, AT and JBV wrote the paper.

12 ACKNOWLEDGMENTS

We are grateful to Julia Ayumi Bopp for input and feedback during the coding and writing, as well as the reviewers for their encouraging comments.

REFERENCES

[1] * Vero Vanden Abeele, Jan Wouters, Pol Ghesquière, Ann Goeleven, and Luc Geurts. 2015. Game-based Assessment of Psycho-acoustic Thresholds. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793132

[2] * Andrea Abney, Brooke White, Jeremy Glick, Andre Bermudez, Paul Breckow, Jason Yow, Rayna Tillinghast-Trickett, and Paul Heath. 2014. Evaluation of recording methods for user test sessions on mobile devices. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658704

[3] ACM SIGCHI. 2020. CHI PLAY 2020 – November 1 – 4, 2020 Ottawa, Canada. Retrieved 2020-04-16 from https://chiplay.acm.org/2020/

[4] Balazs Aczel, Barnabas Szaszi, Alexandra Sarafoglou, Zoltan Kekecs, Šimon Kucharský, Daniel Benjamin, Christopher D Chambers, Agneta Fisher, Andrew Gelman, Morton A Gernsbacher, John P Ioannidis, Eric Johnson, Kai Jonas, Stavroula Kousta, Scott O Lilienfeld, D Stephen Lindsay, Candice C Morey, Marcus Monafò, Benjamin R Newell, Harold Pashler, David R Shanks, Daniel J Simons, Jelte M Wicherts, Dolores Albarracin, Nicole D Anderson, John Antonakis, Hal R Arkes, Mitja D Back, George C Banks, Christopher Beevers, Andrew A Bennett, Wiebke Bleidorn, Ty W Boyer, Cristina Cacciari, Alice S Carter, Joseph Cesario, Charles Clifton, Ronán M Conroy, Mike Cortese, Fiammetta Cosci, Nelson Cowan, Jarret Crawford, Eveline A Crone, John Curtin, Randall Engle, Simon Farrell, Pasco Fearon, Mark Fichman, Willem Frankenhuis, Alexandra M Freund, M Gareth Gaskell, Roger Giner-Sorolla, Don P Green, Robert L Greene, Lisa L Harlow, Fernando Hoces de la Guardia, Derek Isaacowitz, Janet Kolodner, Debra Lieberman, Gordon D Logan, Wendy B Mendes, Lea Moersdorf, Brendan Nyhan, Jeffrey Pollack, Christopher Sullivan, Simine Vazire, and Eric Jan Wagenmakers. 2019. A consensus-based transparency checklist. *Nature Human Behaviour* (dec 2019). https://doi.org/10.1038/s41562-019-0772-6

[5] * Dmitry Alexandrovsky, Maximilian Achim Friehs, Max V. Birk, Rowan K. Yates, and Regan L. Mandryk. 2019. Game Dynamics that Support Snacking, not Feasting. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347151

[6] * Sarah AlSulaiman and Michael S. Horn. 2015. Peter the Fashionista?. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793127

[7] * Maximilian Altmeyer, Pascal Lessel, Marc Schubhan, Vladislav Hnatovskiy, and Antonio Krüger. 2019. Germ Destroyer - A Gamified System to Increase the Hand Washing Duration in Shared Bathrooms. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347157

[8] Craig A. Anderson. 2004. An update on the effects of playing violent video games. *Journal of Adolescence* 27, 1 (feb 2004), 113–122. https://doi.org/10.1016/j.adolescence.2003.10.009

[9] * Dennis Ang and Alex Mitchell. 2017. Comparing Effects of Dynamic Difficulty Adjustment Systems on Video Game Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116623

[10] * Dennis Ang and Alex Mitchell. 2019. Representation and Frequency of Player Choice in Player-Oriented Dynamic Difficulty Adjustment Systems. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347165

[11] * Ivon Arroyo, Matthew Micciollo, Jonathan Casano, Erin Ottmar, Taylyn Hulse, and Ma. Mercedes Rodrigo. 2017. Wearable Learning. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116637

[12] * Louise Ashbarry, Benjamin Geelan, Kristy de Salas, and Ian Lewis. 2016. Blood and Violence. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968111

[13] * Jeremy B. Badler and Alessandro Canossa. 2015. Anticipatory Gaze Shifts during Navigation in a Naturalistic Virtual Environment. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793136

[14] * Alexander Baldwin, Daniel Johnson, and Peta Wyeth. 2016. Crowd-Pleaser. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968100

[15] * Gabriel Barata, Sandra Gama, Joaquim A.P. Jorge, and Daniel J.V. Gonçalves. 2014. Relating gaming habits with student performance in a gamified learning experience. In *Proceedings of the first ACM SIGCHI annual symposium*

on Computer-human interaction in play - CHI PLAY '14. ACM Press. https://doi.org/10.1145/2658537.2658692

[16] Ralf Bender and Stefan Lange. 2001. Adjusting for multiple testing - When and how? *Journal of Clinical Epidemiology* 54, 4 (2001), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

[17] * Max V. Birk, Maximilian A. Friehs, and Regan L. Mandryk. 2017. Age-Based Preferences and Player Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116608

[18] * Max V. Birk, Regan L. Mandryk, and Cheralyn Atkins. 2016. The Motivational Push of Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968091

[19] * Max V. Birk, Regan L. Mandryk, Matthew K. Miller, and Kathrin M. Gerling. 2015. How Self-Esteem Shapes our Interactions with Play Technologies. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793111

[20] * Marion Boberg, Evangelos Karapanos, Jussi Holopainen, and Andrés Lucero. 2015. PLEXQ. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793124

[21] * Jason T. Bowey and Regan L. Mandryk. 2017. Those are not the Stories you are Looking For. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116636

[22] * Evren Bozgeyikli, Andrew Raij, Srinivas Katkoori, and Rajiv Dubey. 2016. Point & Teleport Locomotion Technique for Virtual Reality. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968105

[23] Florian Brühlmann, Serge Petralito, Lena F Aeschbach, and Klaus Opwis. 2020. The Quality of Data Collected Online: An Investigation of Careless Responding in a Crowdsourced Sample. *Methods in Psychology* (2020), 100022. https://doi.org/10.1016/j.metip.2020.100022

[24] Florian Brühlmann and Gian Marco Schmid. 2015. How to measure the game experience? Analysis of the factor structure of two questionnaires. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 18. Association for Computing Machinery, New York, New York, USA, 1181–1186. https://doi.org/10.1145/2702613.2732831

[25] * Jie Cai, Donghee Yvette Wohn, and Guo Freeman. 2019. Who Purchases and Why?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347196

[26] Paul Cairns. 2007. HCI.. not as it should be: Inferential statistics in HCI research. *People and Computers XXI HCI.But Not as We Know It - Proceedings of HCI 2007: The 21st British HCI Group Annual Conference* 1 (2007), 195–201. https://doi.org/10.14236/ewic/hci2007.20

[27] Paul Cairns. 2019. *Doing Better Statistics in Human-Computer Interaction*. Cambridge University Press. https://doi.org/10.1017/9781108685139

[28] * Murat Perit Cakir, Nur Akkuş Çakir, Hasan Ayaz, and Frank J. Lee. 2015. An Optical Brain Imaging Study on the Improvements in Mathematical Fluency from Game-based Learning. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793133

[29] * Jared E. Cechanowicz, Carl Gutwin, Scott Bateman, Regan Mandryk, and Ian Stavness. 2014. Improving player balancing in racing games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658701

[30] * Anjana Chatta, Tyler Hurst, Gayani Samaraweera, Rongkai Guo, and John Quarles. 2015. Get off the Couch. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793115

[31] * Jinghui Cheng, Dorian Anderson, Cynthia Putnam, and Jin Guo. 2017. Leveraging Design Patterns to Support Designer-Therapist Collaboration When Ideating Brain Injury Therapy Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116600

[32] * Sebastian Cmentowski, Andrey Krekhov, and Jens Krüger. 2019. Outstanding: A Multi-Perspective Travel Approach for Virtual Reality Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347183

[33] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK no more: On the preregistration of chi experiments. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3173715

[34] Geoff Cumming. 2009. Dance p 3 Mar09. https://www.youtube.com/watch?v=ez4DgdurRPg

[35] Geoff Cumming. 2013. *Understanding The New Statistics*. https://doi.org/10.4324/9780203807002

[36] Geoff Cumming and Sue Finch. 2005. Inference by eye confidence intervals and how to read pictures of data. *American Psychologist* 60, 2 (2005), 170–180. https://doi.org/10.1037/0003-066X.60.2.170

[37] * Martin Dechant, Ian Stavness, Aristides Mairena, and Regan L. Mandryk. 2018. Empirical Evaluation of Hybrid Gaze-Controller Selection Techniques in a Gaming Context. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242699

[38] * Alena Denisova and Paul Cairns. 2015. The Placebo Effect in Digital Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793109

[39] * Alena Denisova and Eliott Cook. 2019. Power-Ups in Digital Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347173

[40] * Alena Denisova, A. Imran Nordin, and Paul Cairns. 2016. The Convergence of Player Experience Questionnaires. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968095

[41] * Ansgar E. Depping, Colby Johanson, and Regan L. Mandryk. 2018. Designing for Friendship. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242702

[42] * Ansgar E. Depping and Regan L. Mandryk. 2017. Cooperation and Interdependence. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116639

[43] * Ansgar E. Depping, Regan L. Mandryk, Colby Johanson, Jason T. Bowey, and Shelby C. Thomson. 2016. Trust Me. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968097

[44] * Arindam Dey, Hao Chen, Mark Billinghurst, and Robert W. Lindeman. 2018. Effects of Manipulating Physiological Feedback in Immersive Virtual Environments. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242676

[45] Zoltan Dienes. 2008. *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

[46] Zoltan Dienes. 2014. Using Bayes to get the most out of non-significant results. *Frontiers in Psychology* 5 (jul 2014), 781. https://doi.org/10.3389/fpsyg.2014.00781

[47] * Gabriella Dodero, Rosella Gennari, Alessandra Melonio, and Santina Torello. 2014. Towards tangible gamified co-design at school. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658688

[48] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. 291–330. https://doi.org/10.1007/978-3-319-26633-6_13 arXiv:arXiv:1011.1669v3

[49] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the transparency of research papers with explorable multiverse analyses. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–15. https://doi.org/10.1145/3290605.3300295

[50] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. Touchstone2: An Interactive Environment for Exploring Trade-offs in HCI Experiment Design. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–11. https://doi.org/10.1145/3290605.3300447

[51] Malte Elson and Thorsten Quandt. 2016. Digital games in laboratory experiments: Controlling a complex stimulus through modding. *Psychology of Popular Media Culture* 5, 1 (jan 2016), 52–65. https://doi.org/10.1037/ppm0000033

[52] * Katharina Emmerich and Maic Masuch. 2016. The Influence of Virtual Agents on Player Experience and Performance. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968092

[53] * Katharina Emmerich and Maic Masuch. 2017. The Impact of Game Patterns on Player Experience and Social Interaction in Co-Located Multiplayer Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116606

[54] * Katharina Emmerich, Patrizia Ring, and Maic Masuch. 2018. I'm Glad You Are on My Side. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242709

[55] * Zachary Fitz-Walter, Peta Wyeth, Dian Tjondronegoro, and Daniel Johnson. 2014. Exploring the effect of achievements on students attending university orientation. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658700

[56] Wolfgang Forstmeier, Eric Jan Wagenmakers, and Timothy H. Parker. 2017. Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews* 92, 4 (2017), 1941–1968. https://doi.org/10.1111/brv.12315

[57] * Julian Frommel, Kim Fahlbusch, Julia Brich, and Michael Weber. 2017. The Effects of Context-Sensitive Tutorials in Virtual Reality Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116610

[58] * Julian Frommel, Fabian Fischbach, Katja Rogers, and Michael Weber. 2018. Emotion-based Dynamic Difficulty Adjustment Using Parameterized Difficulty and Self-Reports of Emotion. In *Proceedings of the 2018 Annual Symposium on*

Computer-Human Interaction in Play. ACM Press. https://doi.org/10.1145/3242671.3242682

[59] * Julian Frommel, Katja Rogers, Thomas Dreja, Julian Winterfeldt, Christian Hunger, Maximilian Bär, and Michael Weber. 2016. 2084 – Safe New World. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968087

[60] * Julian Frommel, Michael Weber, Katja Rogers, Julia Brich, Daniel Besserer, Leonard Bradatsch, Isabel Ortinau, Ramona Schabenberger, Valentin Riemer, and Claudia Schrader. 2015. Integrated Questionnaires. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793130

[61] * Yue Gao, Kathrin M. Gerling, Regan L. Mandryk, and Kevin G. Stanley. 2014. Decreasing sedentary behaviours in pre-adolescents using casual exergames at school. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658693

[62] * Luc Geurts, Vero Vanden Abeele, Kevin Van Keer, and Ruben Isenborghs. 2014. Playfully learning visual perspective taking skills with sifteo cubes. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658706

[63] * Thomas A. Goldman, Frank J. Lee, and Jichen Zhu. 2014. Using video games to facilitate understanding of attention deficit hyperactivity disorder. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658707

[64] * Antti Granqvist, Tapio Takala, Jari Takatalo, and Perttu Hämäläinen. 2018. Exaggeration of Avatar Flexibility in Virtual Reality. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242694

[65] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 4 (apr 2016), 337–350. https://doi.org/10.1007/s10654-016-0149-3

[66] * Nathan Navarro Griffin, James Liu, and Eelke Folmer. 2018. Evaluation of Handsbusy vs Handsfree Virtual Locomotion. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242707

[67] * Carl Gutwin, Rodrigo Vicencio-Moreira, and Regan L. Mandryk. 2016. Does Helping Hurt?. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968101

[68] * Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to Consider for Tailored Gamification. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347167

[69] * Erik Harpstead, Thomas Zimmermann, Nachiappan Nagapan, Jose J. Guajardo, Ryan Cooper, Tyson Solberg, and Dan Greenawalt. 2015. What Drives People. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793114

[70] * John Harris, Mark Hancock, and Stacey D. Scott. 2016. Leveraging Asymmetries in Multiplayer Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968113

[71] * Jennefer Hart, Ioanna Iacovides, Anne Adams, Manuel Oliveira, and Maria Margoudi. 2017. Understanding Engagement within the Context of a Safety Critical Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116633

[72] * Kieran Hicks, Kathrin Gerling, Patrick Dickinson, and Vero Vanden Abeele. 2019. Juicy Game Design. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347171

[73] Joseph Hilgard, Christopher R. Engelhardt, Bruce D. Bartholow, and Jeffrey N. Rouder. 2017. How much evidence is p >.05? Stimulus pre-testing and null primary outcomes in violent video games research. *Psychology of Popular Media Culture* 6, 4 (oct 2017), 361–380. https://doi.org/10.1037/ppm0000102

[74] * Britton Horn, Seth Cooper, and Sebastian Deterding. 2017. Adapting Cognitive Task Analysis to Elicit the Skill Chain of a Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116640

[75] Kasper Hornbæk. 2011. Some whys and hows of experiments in human-computer interaction. *Foundations and Trends in Human-Computer Interaction* 5, 4 (2011), 299–373. https://doi.org/10.1561/1100000043

[76] Kasper Hornbæk, Søren S. Sander, Javier Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? on the extent and content of replications in human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 3523–3532. https://doi.org/10.1145/2556288.2557004

[77] * Ioanna Iacovides, Anna Cox, Richard Kennedy, Paul Cairns, and Charlene Jennett. 2015. Removing the HUD. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793120

[78] * John Porter III, Matthew Boyer, and Andrew Robb. 2018. Guidelines on Successfully Porting Non-Immersive Games to Virtual Reality. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242677

[79] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2, 8 (aug 2005), e124. https://doi.org/10.1371/journal.pmed.0020124

[80] * Aliya Iskenderova, Florian Weidner, and Wolfgang Broll. 2017. Drunk Virtual Reality Gaming. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116618

[81] * Colby Johanson, Carl Gutwin, Jason T. Bowey, and Regan L. Mandryk. 2019. Press Pause when you Play. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347195

[82] * Colby Johanson, Carl Gutwin, and Regan L. Mandryk. 2017. The Effects of Navigation Assistance on Spatial Learning and Performance in a 3D Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116602

[83] Leslie K. John, George Loewenstein, and Drazen Prelec. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* 23, 5 (2012), 524–532. https://doi.org/10.1177/0956797611430953

[84] * Daniel Johnson, Christopher Watling, John Gardner, and Lennart E. Nacke. 2014. The edge of glory. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658694

[85] Daniel Johnson, M. John Gardner, and Ryan Perry. 2018. Validation of two game experience scales: The Player Experience of Need Satisfaction (PENS) and Game Experience Questionnaire (GEQ). *International Journal of Human Computer Studies* 118 (oct 2018), 38–46. https://doi.org/10.1016/j.ijhcs.2018.05.003

[86] * Dennis L. Kappen, Pejman Mirza-Babaei, Jens Johannsmeier, Daniel Buckstein, James Robb, and Lennart E. Nacke. 2014. Engaged by boos and cheers. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658687

[87] * Dennis L. Kappen, Pejman Mirza-Babaei, and Lennart E. Nacke. 2017. Gamification through the Application of Motivational Affordances for Physical Activity Technology. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116604

[88] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. *Conference on Human Factors in Computing Systems - Proceedings* (2012), 1105–1113. https://doi.org/10.1145/2207676.2208557

[89] * Geoff Kaufman, Mary Flanagan, and Gili Freedman. 2019. Not Just for Girls. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347177

[90] Ken Kelley and Joseph R Rausch. 2006. Sample Size Planning for the Standardized Mean Difference : Accuracy in Parameter Estimation Via Narrow Confidence Intervals. 11, 4 (2006), 363–385. https://doi.org/10.1037/1082-989X.11.4.363

[91] Norbert L. Kerr. 1998. HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review* 2, 3 (aug 1998), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

[92] * Mallory Ketcheson, Zi Ye, and T.C. Nicholas Graham. 2015. Designing for Exertion. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793122

[93] * Soomin Kim, Gyuho Lee, Seo young Lee, Sanghyuk Lee, and Joonhwan Lee. 2019. Game or Live Streaming?: Motivation and Social Experience in Live Mobile Quiz Shows. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347187

[94] * Madison Klarkowski, Daniel Johnson, Peta Wyeth, Cody Phillips, and Simon Smith. 2018. Don't Sweat the Small Stuff. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242714

[95] * Boriana Koleva, Peter Tolmie, Patrick Brundell, Steve Benford, and Stefan Rennick Egglestone. 2015. From Front-End to Back-End and Everything In-Between. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793131

[96] * Andrey Krekhov, Sebastian Cmentowski, Katharina Emmerich, and Jens Krüger. 2019. Beyond Human. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347172

[97] * Andrey Krekhov, Sebastian Cmentowski, Katharina Emmerich, Maic Masuch, and Jens Krüger. 2018. GulliVR. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242704

[98] * Andrey Krekhov, Katharina Emmerich, Philipp Bergmann, Sebastian Cmentowski, and Jens Krüger. 2017. Self-Transforming Controllers for Virtual Reality First Person Shooters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116615

[99] * Sven Krome, Jussi Holopainen, and Stefan Greuter. 2017. AutoGym. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116626

[100] John K. Kruschke. 2013. Bayesian estimation supersedes the T test. *Journal of Experimental Psychology: General* 142, 2 (2013), 573–588. https://doi.org/10.1037/a0029177 arXiv:http://dx.doi.org/10.1037/a0029146

[101] Daniël Lakens. 2014. The 20% Statistician: Observed power, and what to do if your editor asks for post-hoc power analyses. Retrieved 2020-01-22 from https://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html

[102] Daniël Lakens. 2016. Improving your statistical inferences - Week 1: Introduction + Frequentist Statistics. Retrieved 2020-03-10 from https://www.coursera.org/learn/statistical-inferences/home/week/1

[103] Daniël Lakens. 2020. The 20% Statistician: What's a family in family-wise error control? Retrieved 2020-03-16 from https://daniellakens.blogspot.com/2020/03/whats-family-in-family-wise-error.html

[104] Daniel Lakens, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A.J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa Debruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G. Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne Laura Van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M.A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q.X. Nio, Gustav Nilsonne, Cilene Lino De Oliveira, Jean Jacques Orban De Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Marcel A.L.M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano, and Rolf A. Zwaan. 2018. Justify your alpha. *Nature Human Behaviour* 2, 3 (2018), 168–171. https://doi.org/10.1038/s41562-018-0311-x

[105] * Matthew Lakier, Lennart E. Nacke, Takeo Igarashi, and Daniel Vogel. 2019. Cross-Car, Multiplayer Games for Semi-Autonomous Driving. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347166

[106] * Nicole Lane and Nathan R. Prestopnik. 2017. Diegetic Connectivity. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116630

[107] * Michael Lankes, Jüergen Hagler, Georgi Kostov, and Jeremiah Diephuis. 2017. Invisible Walls. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116609

[108] * Michael Lankes, Thomas Mirlacher, Stefan Wagner, and Wolfgang Hochleitner. 2014. Whom are you looking for?. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658698

[109] * Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. 2018. Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242683

[110] * Pascal Lessel, Maximilian Altmeyer, and Nicolas Brauner. 2019. Crowdjump: Investigating a Player-Driven Platform Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347168

[111] * Jingya Li, Erik D. van der Spek, Jun Hu, and Loe Feijs. 2019. Turning Your Book into a Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347174

[112] * Michael Long and Carl Gutwin. 2018. Characterizing and Modeling the Effects of Local Latency on Game Performance and Experience. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242678

[113] * Bernhard Maurer, Ilhan Aslan, Martin Wuchse, Katja Neureiter, and Manfred Tscheligi. 2015. Gaze-Based Onlooker Integration. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793126

[114] * Mitchell W. McEwan, Alethea L. Blackler, Daniel M. Johnson, and Peta A. Wyeth. 2014. Natural mapping and intuitive interaction in videogames. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658541

[115] William J McGuire. 1997. Creative hypothesis generating in psychology: Some useful heuristics. *Annual review of psychology* 48, 1 (1997), 1–30. https://doi.org/10.1146/annurev.psych.48.1.1

[116] Tsuyoshi Miyakawa. 2020. No raw data, no science: another possible source of the reproducibility crisis. *Molecular brain* 13, 1 (2020), 24. https://doi.org/10.1186/s13041-020-0552-2

[117] Marcus R. Munafò, Brian A. Nosek, Dorothy V.M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 1–9. https://doi.org/10.1038/s41562-016-0021

[118] * John E. Muñoz, M. Cameirão, S. Bermúdez i Badia, and E. Rubio Gouveia. 2018. Closing the Loop in Exergaming - Health Benefits of Biocybernetic Adaptation in Senior Adults. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242673

[119] * Juliana Nazare, Anneli Hershman, Ivan Sysoev, and Deb Roy. 2017. Bilingual SpeechBlocks. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116616

[120] * Joshua Newn, Eduardo Velloso, Fraser Allison, Yomna Abdelrahman, and Frank Vetere. 2017. Evaluating Real-Time Gaze Representations to Infer Intentions in Competitive Turn-Based Strategy Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116624

[121] Michèle B. Nuijten. 2016. A spellchecker for statistics. , 151–152 pages. http://blogs.lse.ac.uk/impactofsocialsciences/2018/02/28/statcheck-a-spellchecker-for-statistics/

[122] Open Science Framework. 2016. OSF | Templates of OSF Registration Forms. https://osf.io/zab38/{#}!

[123] * Pablo Ortiz and D. Fox Harrell. 2018. Enabling Critical Self-Reflection through Roleplay with Chimeria. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242687

[124] * Raul Paradeda, Maria José Ferreira, Raquel Oliveira, Carlos Martinho, and Ana Paiva. 2019. The Role of Assertiveness in a Storytelling Game with Persuasive Robotic Non-Player Characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347162

[125] * Pratheep Kumar Paranthaman and Seth Cooper. 2019. ARAPID: Towards Integrating Crowdsourced Playtesting into the Game Development Environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347163

[126] * Taiwoo Park, Tianyu Hu, and Jina Huh. 2016. Plant-based Games for Anxiety Reduction. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968094

[127] * Cale J. Passmore and Regan Mandryk. 2018. An About Face: Diverse Representation in Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 365–380. https://doi.org/10.1145/3242671.3242711

[128] * Johannes Pfau, Jan David Smeddinck, and Rainer Malaka. 2018. Towards Deep Player Behavior Models in MMORPGs. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 381–392. https://doi.org/10.1145/3242671.3242706

[129] * Cody Phillips, Daniel Johnson, Madison Klarkowski, Melanie Jade White, and Leanne Hides. 2018. The Impact of Rewards and Trait Reward Responsiveness on Player Motivation. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242713

[130] Laura Pritschet, Derek Powell, and Zachary Horne. 2016. Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science* 27, 7 (2016), 1036–1042. https://doi.org/10.1177/0956797616645672

[131] Xiaoying Pu, Matthew Kay, Licheng Zhu, and Frederick Conrad. 2019. Designing for preregistration: A user-centered perspective. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–6. https://doi.org/10.1145/3290607.3312862

[132] Daniel S. Quintana. 2019. Synthetic datasets: A non-technical primer for the behavioural sciences to promote reproducibility and hypothesis-generation. (2019). https://doi.org/10.31234/osf.io/dmfb3

[133] * George E. Raptis, Christos A. Fidas, and Nikolaos M. Avouris. 2016. Do Field Dependence-Independence Differences of Game Players Affect Performance and Behaviour in Cultural Heritage Games?. In *Proceedings of the 2016 Annual*

Symposium on Computer-Human Interaction in Play - CHI PLAY '16. ACM Press. https://doi.org/10.1145/2967934.2968107

[134] * Anke V. Reinschluessel and Regan L. Mandryk. 2016. Using Positive or Negative Reinforcement in Neurofeedback Games for Training Self-Regulation. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968085

[135] David C. Rheinheimer and Douglas A. Penfield. 2001. The Effects of Type I Error Rate and Power of the ANCOVA F Test and Selected Alternatives Under Nonnormality and Variance Heterogeneity. *The Journal of Experimental Education* 69, 4 (jan 2001), 373–391. https://doi.org/10.1080/00220970109599493

[136] * Katja Rogers, Matthias Jörg, and Michael Weber. 2019. Effects of Background Music on Risk-Taking and General Player Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347158

[137] Robert Rosenthal. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86, 3 (1979), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

[138] * Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. 2019. Notification in VR. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347190

[139] * Pejman Sajjadi, Edgar Omar Cebolledo Gutierrez, Sandra Trullemans, and Olga De Troyer. 2014. Maze commander. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658690

[140] * Cheryl Savery and Nicholas Graham. 2014. Reducing the negative effects of inconsistencies in networked games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658539

[141] * Mike Schaekermann, Giovanni Ribeiro, Guenter Wallner, Simone Kriglstein, Daniel Johnson, Anders Drachen, Rafet Sifa, and Lennart E. Nacke. 2017. Curiously Motivated. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116603

[142] Ulrich Schimmack. 2015. Questionable Research Practices: Definition, Detect, and Recommendations for Better Practices. Retrieved 2020-01-22 from https://replicationindex.com/2015/01/24/questionable-research-practices-definition-detect-and-recommendations-for-better-practices/

[143] Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. Is It Really Robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology* 6, 4 (2010), 147–151. https://doi.org/10.1027/1614-2241/a000016

[144] Felix D. Schönbrodt. 2016. p-Hacker: Train your p-hacking skills! Retrieved 17.04.2020 from http://shinyapps.org/apps/p-hacker/

[145] * Valentin Schwind and Niels Henze. 2018. Gender- and Age-related Differences in Designing the Characteristics of Stereotypical Virtual Faces. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242692

[146] * Valentin Schwind, Pascal Knierim, Lewis Chuang, and Niels Henze. 2017. "Where's Pinky?". In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116596

[147] * Valentin Schwind, Sven Mayer, Alexandre Comeau-Vermeersch, Robin Schweigert, and Niels Henze. 2018. Up to the Finger Tip. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242675

[148] * Valentin Schwind, Katrin Wolf, Niels Henze, and Oliver Korn. 2015. Determining the Characteristics of Preferred Virtual Faces Using an Avatar Generator. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793116

[149] * Sven Seele, Sebastian Misztal, Helmut Buhler, Rainer Herpers, and Jonas Schild. 2017. Here's Looking At You Anyway!. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116619

[150] * Hanieh Shakeri, Samarth Singhal, Rui Pan, Carman Neustaedter, and Anthony Tang. 2017. Escaping Together. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116601

[151] Larissa Shamseer, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A. Stewart, Douglas G. Altman, Alison Booth, An Wen Chan, Stephanie Chang, Tammy Clifford, Kay Dickersin, Matthias Egger, Peter C. Gøtzsche, Jeremy M. Grimshaw, Trish Groves, Mark Helfand, Julian Higgins, Toby Lasserson, Joseph Lau, Kathleen Lohr, Jessie McGowan, Cynthia Mulrow, Melissa Norton, Matthew Page, Margaret Sampson, Holger Schünemann, Iveta Simera, William Summerskill, Jennifer Tetzlaff, Thomas A. Trikalinos, David Tovey, Lucy Turner, and Evelyn Whitlock. 2015. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. *BMJ (Online)* 349, December 2014 (2015), 1–25. https://doi.org/10.1136/bmj.g7647

[152] * Hitesh Nidhi Sharma, Z. O. Toups, Igor Dolgov, Andruid Kerne, and Ajit Jain. 2016. Evaluating Display Modalities Using a Mixed Reality Game. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968090

[153] * Mike Sheinin and Carl Gutwin. 2015. Quantifying Individual Differences, Skill Development, and Fatigue Effects in Small-Scale Exertion Interfaces. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793129

[154] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology. *Psychological Science* 22, 11 (nov 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[155] * Kristin Siu and Mark O. Riedl. 2016. Reward Systems in Human Computation Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968083

[156] * Milad Soroush, Mark Hancock, and Vanessa K. Bohns. 2018. Investigating Game Mechanics that Target Players' Self-Control While Maintaining Engagement. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242698

[157] * Sharon T. Steinemann, Elisa D. Mekler, and Klaus Opwis. 2015. Increasing Donating Behavior Through a Game for Change. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793125

[158] Denes Szucs and John P.A. Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: A reassessment. https://doi.org/10.3389/fnhum.2017.00390

[159] * Gustavo F. Tondello, Alberto Mora, and Lennart E. Nacke. 2017. Elements of Gameful Design Emerging from User Preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116627

[160] * Gustavo F. Tondello and Lennart E. Nacke. 2019. Player Characteristics and Video Game Preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347185

[161] * Gustavo F. Tondello, Rina R. Wehbe, Rita Orji, Giovanni Ribeiro, and Lennart E. Nacke. 2017. A Framework and Taxonomy of Videogame Playing Preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116629

[162] * Z. O. Toups, Nicole K. Crenshaw, Rina R. Wehbe, Gustavo F. Tondello, and Lennart E. Nacke. 2016. "The Collecting Itself Feels Good". In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968088

[163] * Olivier Tremblay-Savard, Alexander Butyaev, and Jérôme Waldispühl. 2016. Collaborative Solving in a Human Computing Game Using a Market, Skills and Challenges. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968104

[164] * April Tyack, Peta Wyeth, and Daniel Johnson. 2016. The Appeal of MOBA Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968098

[165] April Tyack and Elisa D Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. *CHI 2020 - Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://doi.org/10.1145/3313831.3376723

[166] April Tyack, Peta Wyeth, and Madison Klarkowski. 2018. Video game selection procedures for experimental research. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. Association for Computing Machinery, New York, New York, USA, 1–9. https://doi.org/10.1145/3173574.3173760

[167] * Kellie Vella, Daniel Johnson, and Leanne Hides. 2015. Playing Alone, Playing With Others. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793118

[168] * Kellie Vella, Christopher James Koren, and Daniel Johnson. 2017. The Impact of Agency and Familiarity in Cooperative Multiplayer Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116622

[169] Kellie Vella, Madison Klarkowski, Daniel Johnson, Leanne Hides, and Peta Wyeth. 2016. The social context of video game play: Challenges and strategies. In *DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems: Fuse*. Association for Computing Machinery, Inc, New York, New York, USA, 761–772. https://doi.org/10.1145/2901790.2901823

[170] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts : Results of a Self-Reported Survey. (2020). https://doi.org/10.31219/osf.io/3bu6t

[171] * Guenter Wallner. 2015. Sequential Analysis of Player Behavior. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793112

[172] * Guenter Wallner and Simone Kriglstein. 2016. Visualizations for Retrospective Analysis of Battles in Team-based Combat Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968093

[173] * Justin D. Weisz, Maryam Ashoori, and Zahra Ashktorab. 2018. Entanglion. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242696

[174] * Lindsay Wells, Aran Cauchi-Saunders, Ian Lewis, Lorenzo Monsif, Benjamin Geelan, and Kristy de Salas. 2016. Mining for Gold (and Platinum). In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968112

[175] * Matthew Alexander Whitby, Sebastian Deterding, and Ioanna Iacovides. 2019. "One of the baddies all along". In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347192

[176] * Laura A. Whitlock, Anne Collins McLaughlin, William Leidheiser, Maribeth Gandy, and Jason C. Allaire. 2014. Know before you go. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play - CHI PLAY '14*. ACM Press. https://doi.org/10.1145/2658537.2658703

[177] Jelte M. Wicherts, Coosje L.S. Veldkamp, Hilde E.M. Augusteijn, Marjan Bakker, Robbie C.M. van Aert, and Marcel A.L.M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology* 7, NOV (2016), 1–12. https://doi.org/10.3389/fpsyg.2016.01832

[178] * Graham Wilson and Mark McGill. 2018. Violent Video Games in Virtual Reality. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM Press. https://doi.org/10.1145/3242671.3242684

[179] Jacob O. Wobbrock and Julie A Kientz. 2016. Research contribution in human-computer interaction. *interactions* 23, 3 (apr 2016), 38–44. https://doi.org/10.1145/2907069

[180] * Donghee Yvette Wohn, Peter Jough, Peter Eskander, John Scott Siri, Masaho Shimobayashi, and Pradnya Desai. 2019. Understanding Digital Patronage. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347160

[181] * Priscilla N.Y. Wong, Jacob M. Rigby, and Duncan P. Brumby. 2017. Game & Watch. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116613

[182] * Daniel Yule, Bonnie MacKay, and Derek Reilly. 2015. Operation Citadel. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793135

[183] * Anna Zamansky, Dirk van der Linden, Sofya Baskin, and Vitaliya Kononova. 2017. Is My Dog "Playing" Tablet Games?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17*. ACM Press. https://doi.org/10.1145/3116595.3116634

[184] * Majed Al Zayer, Sam Tregillus, and Eelke Folmer. 2016. PAWdio. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM Press. https://doi.org/10.1145/2967934.2968079

[185] * David Zendle, Paul Cairns, and Daniel Kudenko. 2015. Higher Graphical Fidelity Decreases Players' Access to Aggressive Concepts in Violent Video Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. ACM Press. https://doi.org/10.1145/2793107.2793113

[186] * Hao Zhang, Qiong Wu, Chunyan Miao, Zhiqi Shen, and Cyril Leung. 2019. Towards Age-friendly Exergame Design. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM. https://doi.org/10.1145/3311350.3347191

* Reference included in literature review.