Empirical Research in Affective Computing: An Analysis of Research Practices and Recommendations

Conference Paper · October 2021

DOI: 10.1109/ACII52823.2021.9597418

5 authors, including:

Janet Wessler
Deutsches Forschungszentrum für Künstliche Intelligenz
8 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Tanja Schneeberger
Deutsches Forschungszentrum für Künstliche Intelligenz
21 PUBLICATIONS **70** CITATIONS
SEE PROFILE

Bernhard Hilpert
Deutsches Forschungszentrum für Künstliche Intelligenz
3 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Alexandra Alles
Deutsches Forschungszentrum für Künstliche Intelligenz
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Empirical Research in Affective Computing: An Analysis of Research Practices and Recommendations

Janet Wessler, Tanja Schneeberger, Bernhard Hilpert, Alexandra Alles, Patrick Gebhard
German Research Center for Artificial Intelligence
Saarland Informatics Campus, Saarbrücken, Germany
firstname.lastname@dfki.de

***Abstract*—In the last decade, empirical sciences have faced a tremendous change in the way of conducting research. As a broad interdisciplinary field, research in Affective Computing often employs empirical user studies. The current paper analyzes research practices in Affective Computing and deduces recommendations for improving the quality of methods and reporting. We extracted a total of k = 65 empirical studies from the two most recent International Conferences on Affective Computing & Intelligent Interaction (ACII) '17 and '19. Three raters summarized characteristics of studies (e.g., number of experimental studies) and how much methodological (e.g., participant characteristics) and statistical information (e.g., degrees of freedom) were missing. Also, we conducted a p-curve analysis to test the overall evidential value of findings. Results showed that 1. in at least half of the studies, one important information about statistical results was missing, and 2. those k = 31 studies that had reported all necessary information to be included into the p-curve showed evidential value. In general, all criteria were never met in one single study. We provide concrete recommendations on how to implement open research practices for empirical studies in Affective Computing.**

***Index Terms*—Open Science, Questionable Research Practices, p-Curve Analysis**

I. Introduction

Affective Computing is a broad interdisciplinary field [1]. It covers research topics like interactive virtual agents and human interaction with them, emotion recognition, and computational models of affect. Some of the major conferences that feature Affective Computing research are the International Conference on Affective Computing & Intelligent Interaction (ACII), Intelligent Virtual Agent (IVA) conference, and the Conference on Intelligent User Interfaces (IUI). Affective Computing's interdisciplinary approach resulted in fruitful developments and new findings over the years. It continues to be an inspiring venue for psychology, as well as cognitive, physiology and computer sciences [1]. However, this precious union comes with some challenges. Empirical science has undergone tremendous development and change in the past ten years. What had been framed as "replication crisis" in the beginning [2], is now seen as a "renaissance" [3]. Many methodological approaches were questioned from a meta-scientific point of view, recommendations were formulated and, in consequence, applied by journals, editors, reviewers, and authors. One major driver of this change was the Open Science Movement (e.g., [4]). This change in methodology and thinking also spread to other disciplines like biology [5] and physics [6]. Because Affective Computing consists of such a close synthesis between psychology and computer science, one might expect that this field is also affected by changes in psychological science and science in general. There have been first approaches to the handling of questionable research practices in this research area. In 2018, the first *Workshop on Methodology and the Evaluation of Intelligent Virtual Agents* was held. The *second workshop in 2019* had a strong focus on questionable research practices and formulated actionable points to improve empirical research in the field. Therefore, in this paper, we investigate the available methodological information in experimental research papers from the area of Affective Computing and present recommendations on how to improve empirical research and its reporting in a specific field of human-computer interaction. For our analysis, we focus on the two previous ACII conferences (2017, 2019). The current paper aims to examine the current empirical research practices in the field of Affective Computing. We provide an overview of which empirical research is conducted in Affective Computing and how it is reported according to the Open Science Community standards. Moreover, we meta-analyze for systematic biases in publications applying a p-curve analysis.

Our goal is to raise awareness on how methodological standards can be improved and that open science is crucial to strengthen the work of the whole Affective Computing research community. In order to do so, we first analyze existing work based on suitable criteria. Based on this, we present recommendations that support high methodological standards.

II. Background

Trust is a crucial factor in scientific research. Trust among researchers is fundamental for fruitful collaborative relationships and activities [7]. Researchers should trust that research was performed as described in published scientific papers, that relevant information has been disclosed, and that data has not been manipulated [8], [9]. Not only internally between researchers, but also externally in relationships between science and public, for example, granting agencies or tax payers,

This work is partially funded by the German Research Foundation (DFG) within the DEEP project (funding code 392401413).

trust is an essential factor for facilitating interactions [7]. A violation of ethical norms and values held in common by the scientific community damages the integrity of science [9]. The trust in science and scientific findings all of a sudden dwindled at the beginning of the last decade when cases of obvious fraud and data fabrication became public, and reknown researchers had to leave their positions and retract widely cited articles [3]. Apart from such very rare obvious abuses, there are still more subtle ways, called Questionable Research Practices (QRPs), to influence data collection and analysis process, leading to systematic and large biases in the research literature. One reason for applying QRPs could be the implicit rule that mostly significant findings are published [3], [10]. For many decades, scientists relied on the arbitrarily set significance threshold of $p < .05$ [11]. This implies that overall, there should be less than 5% of false positive findings in the literature, in which the null hypothesis is rejected although it is true. During the replication crisis, however, scientists discovered that the rate of false positive findings was detrimentally higher than this [10]. In a huge replication project, $k = 100$ experimental and correlational studies from three major psychological journals published in 2008 were directly replicated by other laboratories with high-powered studies [4]. While 97% of the original studies showed significant effects, only 36% of replication studies did so. Moreover, effect sizes of the replications were, on average, only about half of the sizes of the original effect sizes. This replication project raised the awareness of the importance of transparent reporting and open science methods. A major reason for such rather disappointing numbers is the flexibility researchers have during data collection, statistical analyses, and reporting in the form of QRPs [10].

A. *Questionable Research Practices*

QRPs lay on a continuum between deliberate misconduct (fabrication, falsification, and plagiarism) and responsible conduct of research, whereas the former represents the worst behavior and the latter the ideal behavior [12]. Since no straightforward definition of QRPs exists [13], they fall into an ethical "gray zone" [14] between acceptable and unacceptable [15]. The analysis of the prevalence of QRPs leads to the assumption that applying QRPs is common among empirical research [16]. Therefore, this survey aims to understand that applying QRPs is counterproductive for realizing consistent high science quality [17]. QRPs can lead to the publication of misinterpreted results and are therefore a major threat to any scientific community and maybe even for the whole society [13]. QRP either happen during the research process (e.g., excluding data points based on post hoc criteria, falsifying data, deciding whether to collect more data after looking to see whether the results were significant) or during the writing process (e.g., failing to report all of a study's conditions or dependent measures, "rounding off" a p value) [16]. Computer simulations show that QRPs, especially when combined, can very easily lead to significant results [10]. Explanations for the occurrences of QRPs reach from inadequate training of researchers to the pressures and incentives to publish in certain outlets, and the demands and expectations of journal editors and reviewers [15]. QRPs are especially detrimental for underpowered studies. Statistical power is the likelihood to find an effect that actually exists. To have reliable study results, studies should have a statistical power of at least 80% [18]. Otherwise, the reasons for non-significant results remain unclear. There could be either truly no effect (true negative) or there is a true effect, but the study had not enough power to detect it (false negative). If underpowered studies show significant effects, this is most likely a false positive finding resulting from chance or QRPs when the effect size is small [3]. In the following, the most crucial QRPs are presented.

***p*-Hacking** occurs when researchers conduct data collection or analyses until non-significant results become significant [10], [19], [20]. Several practices lead to p-hacking [19], for example, including or dropping outliers after analyses, choice of covariates, or conducting statistical analyses during the experiment to decide whether or not to continue the data collection [16]. One special case of p-hacking are **failures of reporting**. These include measuring many dependent variables and deciding which to report in the paper, selectively reporting studies that "worked", p–hacking can be a result of the attempt to understand data and is often not a result of a malicious intent [3]. Still, p-hacking has adverse effects on research. By applying p-hacking, any false hypothesis can be turned into one with statistically significant support [3]. Therefore, it increases the probability of publishing a false-positive as a true-positive finding [3], [21]. p-hacking represents a major threat to the integrity of empirical research that relies on hypothesis testing [3], [10]. Nelson et al. (2018) assume that due to p-hacking many findings in the literature are false positives. Researchers may base their theoretical background and hypotheses on these false positive results, which leads to false assumptions for new studies and increases the risk of running studies that afford time and money but can never reveal the expected effects based on the published literature.

HARKing means hypothesizing after results are known [22]. Kerr differentiates two categories of HARKing. The first category refers to the practice of presenting one or more post hoc hypotheses (i.e., developed after data analysis) as having been the scope of the study from the start. In the second category, researchers exclude one or more a priori hypotheses from their research report. Both practices lead to the readers' (e.g., a reviewer) understanding that a larger proportion of the researchers' "a priori" hypotheses are supported. The costs of several types of this flexible contortion of hypotheses to fit the data are interpreted differently [15], [22]–[25]. However, it seems that all discussants see it as a questionable research practice. Some researchers see it as an ethical concern [15], [22], [25]; a violation of a fundamental principle of communicating scientific research honestly and completely [22], as it involves deception [15]. Also, we argue that HARKing has detrimental effects on the credibility and trustworthiness of research results. The reader might assume that a study has followed a deductive method, whereas it did not.

B. Open Research Practices

"Two central values of science are openness and reproducibility." (p. 139, [26]). To meet these values and conduct trustworthy research, researchers can apply several open research practices, which are important for planning and reporting empirical research. They all aim at improving the quality of published research.

Preregistration is one of the major developments of the past decade. The idea is to formulate research questions and a priori hypotheses before data collection, to describe the design, procedure, and measurements in detail, and to plan statistical tests for hypotheses before the start of a study [27]. This practice then reduces the risk for HARKing and *p*-hacking [28] because hypotheses and statistical analyses are clear in advance. Doing this requires more time for planning studies but then saves time after the study is finished. It is also possible but not necessary to formulate exploratory hypotheses. However, exploratory analyses can be conducted in addition to the planned confirmatory analyses in order to maintain flexibility and room for surprising findings. Preregistration has many benefits to science as well as for individual researchers (see [27] for an overview), that can adequately demonstrate that *p*-hacking was not applied [3]. Also, a detailed planning reduces flaws in the study's design or setup.

Registered reports represent a new approach for publishing scientific work. This new approach incorporates preregistration of experimental designs and peer review before data collection [26]. With the reviewers' support, authors can refine the proposed study (or studies) and makes changes in advance. For example, a reviewer might suggest a better operationalization for a construct that the authors could implement in the study. The interesting part here is that the paper will be accepted *before* data collection starts. If the authors follow the preregistered plan, the paper will be published independently of the results, thus reducing the pressure to find a significant result. Registered reports are hence one possibility to enhance the credibility of empirical results in articles without endangering research success [26].

Reporting standards for empirical research exist in many fields [29]–[32]. Because the interdisciplinary field of Affective Computing involves psychology and cognitive science [1], the community can orient on the reporting standards of the American Psychological Association [33]. Their publication manual provides article reporting standards for qualitative, quantitative and mixed methods designs. By applying these unified standards for reporting, research gets more comprehensive, accurate, and transparent for readers.

Open methods and results support improving research practices is to ask authors to practice open science. This includes disclosing all study materials like questionnaires and stimulus material. Moreover, anonymized data and analysis scripts (e.g., SPSS syntax, R code) can be made public. [10]. This procedure limits the possibility for selective reporting of variables and gives other researchers the opportunity for replications. In psychological empirical science, disclosure is not a requirement everywhere yet but seems slowly to become the norm [3]. To acknowledge open science, research articles can be certified with badges for preregistration, open data, and open materials [34].

Replications represent the basic idea of science. A scientific finding should be reproducible by any other researcher. If a finding is replicable, this strengthens the confidence in the stability of the claimed effect and its generalizability to different contexts (e.g., cultures). Moreover, replications are one way to identify which published findings are based on true effects and which could be the result of chance [35]. In reality, scientists have an urge to search for new and fancy findings. However, basic effects should prove reliable in order to build new ideas and findings on them. In the last decade, the amount of conceptual and exact replication studies increased [3]. In general, it is not trivial to decide whether a replication failed or not because different results can emerge for various reasons [3]. To run the same study a second or third time is difficult, for example, due to differences in time and location of the studies [3]. Nevertheless, examples of replications in Affective Computing already exist [36].

III. Methods

The basis of our analysis were the lists of publications of the *ACII'17* and *ACII'19* proceedings provided by IEEE Xplore. Only full papers are considered. Both proceedings together count 204 full papers (96 for ACII'17, 108 for ACII'19). The data used for this paper can be found on *OSF*.

A. Exclusion Criteria

Papers are not considered if they: (a) address a doctoral consortium, (b) not reporting an experimental or correlational empirical study; this includes studies either reporting no empirical study or empirical studies with a secondary goal, for example for building a database or model validations, proof-of-concept without experimental manipulation, pilot study, (c) reporting empirical studies without human subjects (e.g., the participants are robots). See Fig. 1 for an overview.

The remaining papers include studies that report an empirical user study fulfilling the following criteria: (a) the study had an experimental or quasi-experimental research design (at least two groups were compared between or within participants), (b) the study had a correlational research design, (c) the dependent variable was measured on human subjects, and (d) the studies were of primary goal.

At the end, the review considers 59 papers with $k = 65$ studies, with a total sample size of $N = 13807$.

For the *p*-curve analysis, we excluded studies that either did not report all necessary information (i.e., a test statistic with degrees of freedom) or had conducted tests that could not be included in the *p*-curve (e.g., non-parametric tests). These criteria led to 31 statistical values in the *p*-curve analysis.

B. Characteristics of Analyzed Papers

Eight characteristics that describe the empirical work reported in papers are analyzed [33]: *1) Research design* (correlational vs. experimental vs. quasi-experimental vs. mixed): A

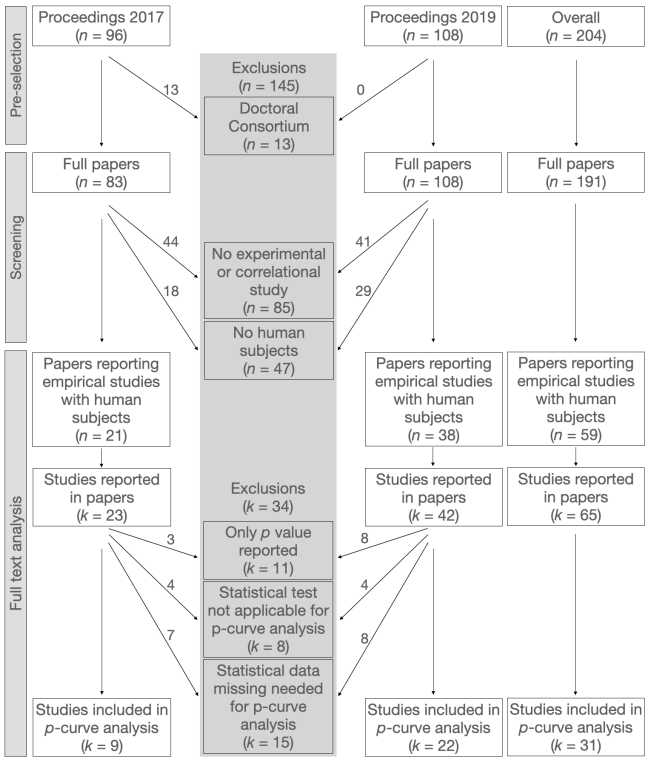

Figure 1: Exclusion flow chart.

research design was correlational if variables were measured but not manipulated, experimental if participants were randomly assigned to different groups and quasi-experimental if groups had not been randomly assigned. *2) Research question* (explorative vs. confirmative vs. mixed vs. none): Studies aimed at exploring a research question were considered exploratory; they were confirmative if authors wrote that they "hypothesize/ assume/ predict / etc."; they were mixed if both kinds of research questions occurred. *3) Experimental design* (within vs. between vs. mixed vs. none vs. unclear): percentages refer to the ratio of designs among all experimental studies, because design only classified for these. *4) Number of independent variables*: The amount of independent variables for experimental studies. *5) Setting* (laboratory vs. field vs. online). *6) Sample size*. *7) Number of reported statistical tests*. *8) Ratio of reported significant versus non-significant tests*: When a result was tested for significance and reported either within text or table, we counted both significant and non-significant results and calculated the ratio.

We analyzed a ninth characteristic — whether studies were preregistered. However, none of the studies claimed to be preregistered, therefore we excluded this characteristic from the reported results.

C. Soundness of Reporting

To evaluate the soundness of reporting, we analyze the methodological and statistical information given for each included study. We oriented on the reporting standards for quantitative research to describe the sample, and statistical results [33]. We chose ten criteria which are key factors for readers to interpret the results of experiments. *Ethical Approval.* We examined if authors reported about any kind of ethical approval with the search terms "ethic", "IRB", "review", and "approval" within each paper. Please note that some studies with human participants are exempt from needing IRB approval (e.g., no identifiable information is collected). However, this should still be mentioned in the paper (i.e., IRB approval or exemption was obtained). *Sample Size Planning* justifies how the sample size was determined. The sample size could be, for example, the result of an a priori power analysis [37] or of a rule of thumb [18]. If such a justification was not given, this information counted as missing. *Sample Size* is the total number of participants in a study. This information is missing when there was no explicit information about it, for example, "N = x", or "data from x participants". *Participant characteristics* counted as missing when there was no information on either participants' age or gender. *Exclusion criteria.* Information counted as missing when (1) the number of participants was smaller in the analyses than initially reported in the sample characteristics and (2) no justification was given for why participants were excluded. *Statistical Analysis.* This information was missing when there was no statement about which kind of statistical analysis was calculated (e.g., ANOVA, *t*-test, regression). *Statistical value (e.g., F, t)*, *degrees of freedom*, *p-value* and *effect size estimate* are part of a good practice test reporting. The information was considered missing when one these values was not given for at least one significant or non-significant result.

D. p-Curve Analysis

p-curve analysis is a meta-analytical statistical method that tests the distribution of *p*-values. The assumption is that when evidential value for claimed effects in certain literature exists, then the distribution of *p*-values $< .05$ should be right skewed. This means the majority of *p*-values should fall to p ¡ .01 rather than between $p = .04$ and $p = .05$. When there is no true effect, however, all *p*-values should evenly distribute between 0 and .05. This is because for a non-existing effect, each *p*-value between 0 and 1 is equally likely to emerge. When there is a systematic bias in the literature, however, the *p*-curve should be left-skewed with the majority of values being closer to .05. Systematic bias could result from *p*-hacking. To get their papers published, researchers might be motivated to extract a *p*-value below .05 from the data. Once they reach this threshold, they stop their efforts and report the results. The advantage of *p*-curve compared to other meta-analytical methods are that (1) only few statistical information is needed to conduct the analysis and (2) the method can be applied to very different contents and research questions in order to check for systematic biases in a certain literature landscape. Thus, the current *p*-curve serves as a summary of the evidential value of studies found in the ACII'17 and '19 conferences. We conducted the *p*-curve analysis using the online *p-curve app*.

IV. Results

A. Characteristics of Review Data Set

In total, our data set consisted of 65 studies, 23 from 2017, 42 from 2019. Table I gives an overview of the studies' characteristics. The mean ratio of reported significant versus non-significant results resided to 59.59% (*SD* = 24.98) for 2017 and 56.45% (*SD* = 31.81) for 2019, respectively.

Table I: Characteristics of reported studies in papers from the ACII'17 and '19 conferences.

Characteristics	2017	2019	Overall
K (Studies)	23	42	65
Research design			
Correlational	1 (4%)	5 (12%)	6 (9%)
Experimental	15 (65%)	34 (81%)	50 (76%)
Quasi-Experimental	7 (30%)	3 (7%)	10 (15%)
Research question			
Explorative	12 (52%)	9 (21%)	21 (32%)
Confirmative	9 (39%)	27 (64%)	37 (56%)
Mixed	2 (9%)	4 (10%)	6 (9%)
None	-	2 (5%)	2 (3%)
Experimental design			
within	15 (65%)	10 (24%)	26 (39%)
between	1 (4%)	13 (31%)	14 (21%)
mixed	6 (26%)	13 (31%)	19 (29%)
none	1 (4%)	5 (12%)	6 (9%)
unclear	-	1 (2%)	1 (2%)
Number of independent variables			
Min	1	1	1
Max	4	5	5
M	2.20	2.21	2.21
SD	1.11	1.16	1.14
Setting			
Laboratory	12 (52%)	31 (74%)	44 (67%)
Field	3 (13%)	6 (14%)	9 (14%)
Online	7 (30%)	5 (12%)	12 (18%)
Unclear	1 (4%)	-	1 (2%)
Sample size			
Min	14	12	12
Max	5057	716	5057
M	381.77	128.76	215.73
SD	1105.98	174.48	664.99
Number of reported statistical tests			
Min	2	1	1
Max	141	220	220
M	31.42	25.66	27.46
SD	38.98	37.45	37.70

Note. Absolute numbers of studies and percentage in parentheses. Due to rounding of the percentages, they do not necessarily sum up to 100%.

There was a wide variation in sample size in both conferences, from $N = 12$ up to $N > 5000$. The majority were laboratory studies which applied a within-participants and an experimental design. Also, we observed a wide variation in the number of reported statistical tests, from 1 to 220.

B. Soundness of Reporting

We analysed all papers regarding the soundness of reporting examining ten methodological and statistical criteria (Tab. II).

Table II: Percentage of studies not reporting information.

Missing information	2017	2019	Overall
Ethical approval	19 (83%)	25 (60%)	44 (68%)
Sample size planning	23 (100%)	41 (98%)	64 (98%)
Sample size	1 (4%)	0 (0%)	1 (2%)
Participant characteristics	10 (44%)	12 (29%)	22 (34%)
Exclusion criteria	0 (0%)	7 (17%)	7 (11%)
Statistical analysis	2 (9%)	4 (10%)	6 (9%)
Test statistic (e.g., F, t)	13 (57%)	24 (57%)	37 (57%)
Degrees of Freedom	18 (78%)	35 (83%)	53 (82%)
p-value	10 (44%)	24 (57%)	34 (52%)
Effect size estimate	21 (91%)	36 (86%)	57 (87%)

Note. Absolute numbers of studies and percentage in parentheses.

C. p-Curve Analysis

For both conferences, $k = 31$ studies had reported all statistical information necessary to conduct a *p*-curve analysis, that is a test statistic with correct degrees of freedom. Of the 31 values, 16 *p*-values had been reported correctly in the original papers, 11 *p*-values had rounding errors (e.g., $p = .00019$ calculated from *p*-curve, but $p < .01$ reported in the paper; this should be $p < .001$; $p = .01164$ calculated form *p*-curve, but $p = 0.011$ reported in the paper; this should be $p = 0.012$). Four values were reported incorrectly, which led to two cases which had originally been reported as significant, but emerged as non-significant in the *p*-curve analysis (e.g. $p = .05319$ calculated form *p*-curve, but $p = 0.0471$ reported in the paper). Thus, the *p*-curve analysis uses 29 significant values (Fig. 2).

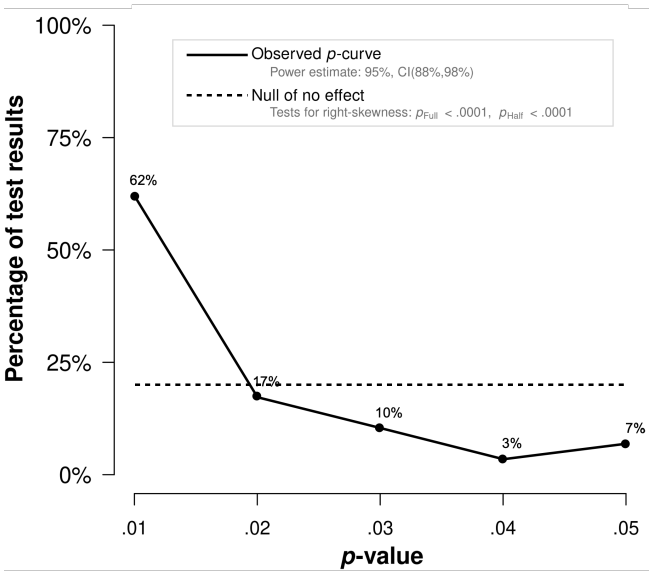

Figure 2: *p*-curve.

The *p*-curve shows that the studies contain evidential value, because the *p*-curve is right-skewed [38], [39], $p = .0003$. Most of the *p*-values (62%) thus are $p < .01$. This shape indicates that most likely these *p*-values were not affected by *p*-hacking. The estimated power of tests included in the *p*-curve is 95%, 90% CI (88%, 98%), which is satisfying.

V. Discussion, Limitations and Future Work

The current paper examined empirical research presented on ACII'17 and '19. We analyzed the characteristics of the empirical studies, their soundness of reporting and conducted a *p*-curve analysis to test their evidential value. By systematically classifying the papers, we discovered that important

information is missing, necessary to interpret and evaluate the findings. However, the studies included in the *p*-curve analysis showed evidential value. Building on these results, we will provide recommendations for future work in the field.

In summary, three key points arise from the results. First, the number of empirical studies including human subjects in Affective Computing is rising. The studies eligible to our analysis (experimental or correlational set-up, DV measured in human subjects) almost doubled from 23 in 2017 to 42 in 2019 — even though the number of studies in the proceedings only rose slightly (from 96 to 108). This demonstrates that empirical studies play an important part in the Affective Computing literature. Second, clear reporting standards and sufficient information to evaluate and interpret findings are missing. The soundness of statistical result reporting did not differ greatly between 2017 and 2019. Especially criteria like reporting of ethical approval, participant characteristics and reports of statistical results (test statistic, degrees of freedom, *p*-value, and effect size) were reported inconsistently across studies. For example, a considerable amount of studies failed to report participant characteristics, although this proportion decreased from 44% in 2017 to 29% in 2019. Demographic information about the sample - especially age and gender - is easily accessible and yields important information on how to generalize study results. Although studies were very consistently reporting the size of the sample, they almost all failed to report how this size had been determined. This information is necessary to assess how likely a study is to find an existing effect, that is to make inferences about statistical power. Exclusion criteria did not seem missing throughout the studies. However, this information was only considered missing if the study description initially reported a sample size that differed from the final sample size used to conduct the analysis. It is not clear if authors had excluded participants without reporting it. The number of reported statistical tests varied greatly from 1 reported result up to 220. A high number of statistical tests increases the likelihood for false positives. Moreover, test statistics, degrees of freedom, *p*-value, and effect size were missing more than half of the time at least for one reported result — degrees of freedom were even missing in 82% of studies at least once. Such a non-transparent reporting practice decreases the trustworthiness of findings [8], [9]. Readers cannot know for why authors are holding back information and could thus be suspicious of QRPs. Such practices furthermore reduce the comparability of findings, the possibility for meta-analyses and for using effect sizes as an orientation for sample size planning.

Third, about half of the studies (48%) had reported all statistical information necessary to be included in the *p*-curve analysis. For this selection, the *p*-curve analysis showed evidential value with a satisfying estimated power of 95% — most *p*-values were $< .01$ and only a few at the critical significance threshold of $p < .05$. It seems that among these studies, *p*-hacking was not of major concern. In line with this conclusion, we observed a significance ratio of 57.52% across all reported *p*-values. This is far from significance ratios up to 97% in psychology [4]. It seems that, in the ACII community, there is no pressure to present only significant results to publish. This is promising and paves the way for even more transparent and clear reporting standards. This is why, we provide recommendations for future standards of reporting on which researchers can orient (Sec. VI).

The current paper analyzed studies from the past two ACII conferences. Three raters divided the work and coded the studies. Unclear cases were discussed between raters and the authors of this paper. Moreover, each rater double-checked a random set of 10% of studies from another rater and discussed inconsistencies. One limitation is that not every rater analyzed all papers. Thus, we could not calculate interrater reliabilities. Future work could,for example, examine the prevalence of and opinion on questionable research practices in a survey among researchers in the field of Affective Computing.

VI. Recommendations

Psychology's renaissance has demonstrated that change in how science is conducted and reported is urgently needed and, more importantly, that it is possible [3]. We now want to take a first step for the field of Affective Computing, providing a set of recommendations for improving research and reporting in future papers, mainly but not limited to papers published on ACII. These recommendations are drawn from several papers on empirical research methods [3], [18]–[21], [26], and additionally build on the results of the current paper. We list 18 recommendations III and describe them in more detail afterward. These recommendations are optional guidelines, and implementing all of them might not always be feasible. However, we recommend starting with small first steps — choosing recommendations that seem feasible and applicable to one's own research.

Prepare your study well. Planning empirical studies takes time, because many aspects have to be considered. We consider obtaining ethical approval, preregistrations and statistical power considerations as important here. Empirical studies should ensure that the rights of participants are protected and that guidelines for research with human subjects are followed. Researchers should thus obtain approval for their studies from their local review board. This is especially important for studies using deception, physiological measures, or new methodological approaches. In a preregistration, the hypotheses, all variables (dependent and independent) and their operationalization, and exclusion criteria are described, and the sample size is planned. Preregistration protocols are available on *OSF* and *AsPredicted*. Here, researchers can timestamp and publish their plans. These can be used in anonymized form for peer-review. One step further than preregistrations are registered reports. These might be especially interesting for biannual conferences, such as ACII. Such conferences could offer the registration of reports in the year in which the conference does not take place. The peer-review process then advances data collection. A hybrid reviewing process could allow conventional submissions as well as the submission of registered reports. Simmons et al. give recommendations

Table III: Recommendations for Affective Computing Community.

1. Obtain ethical approval. Let the local review board check the study.
2. Preregister user studies. Use registration protocols on *OSF* and *AsPredicted*. Explore the possibility of registered reports.
3. Increase statistical power. Plan your sample size for a power of at least 80% (e.g., with G*Power; [37]).
4. Choose an appropriate research design. The design should be suitable to test hypotheses with sufficient power (e.g., within-participants).
5. Check assumptions for statistical tests. Ensure the robustness of your results by inspecting your data (e.g., normality and homoscedasticity for ANOVA).
6. Check for statistical outliers. Outlier criteria should be ideally justified in a preregistration.
7. Report results with and without the covariate. If any, inclusion of covariate should be ideally justified in a preregistration.
8. Reduce the number of statistical tests. Keep the number of tests to a necessary minimum and correct for multiple testing.
9. Follow APA reporting standards. Write study report accordingly.
10. Report all dependent and independent variables. Justify why you focused on specific variable, ideally in a preregistration, but report all manipulated and measured variables in text or in supplemental materials.
11. Justify your sample size. Use tools for a priori sample size planning or rules of thumb.
12. Report recruitment strategies. Make transparent how you recruited your participants (e.g., on campus, via social networks) and which incentives they received.
13. Report participant characteristics. Include detailed information about your sample, at least participants' gender and age (mean and standard deviation).
14. Report exclusion criteria. Describe how you selected participants (e.g., first year Psychology students only). If you excluded participants, justify these exclusions.
15. Report the statistical analysis in detail. Give information on which statistical analysis was performed (e.g., independent samples t-test), the test statistic, degrees of freedom, p-value and an effect size estimate (e.g., Cohen's d for for between-participants comparisons) for both significant and non-significant results, for example, $t(146) = -2.61$, $p = .005$, $d = 0.61$. Provide means and standard deviations for each group.
16. Disclose materials and data. Provide access to the questionnaires, materials, data and analysis code (e.g., SPSS syntax, R code), for example in an project.
17. Provide access to system software. Provide a software container, e.g., *Docker*, with the system files and a documentation.
18. Do and value replications. Run exact or conceptual replications of your own work, or — if feasible — of other laboratories. As a reviewer, value replications as much as novel research.

regarding statistical power [18]. To reach a minimum of 80% power, the sample size should be planned before the data collection and any data analysis. However, this usually requires researchers to know about effect sizes. Especially when researching new topics, effect sizes are often unknown. One heuristic approach is to have at least $n = 50$ participants per cell in order to avoid an underpowered study [18], [40].

Beware of methodology and statistical analyses. Planning and analyzing empirical studies properly is complex. Researchers should therefore have a background on empirical research methods and statistics. Very generally, they should choose appropriate research designs to test hypotheses, choose sufficient sample sizes to increase power, and appropriate statistical methods, ideally with criteria defined in the preregistration [41], [42]. Table III provides further recommendations.

Follow APA reporting standards. The APA manual (7th edition) [33] provides a structure for empirical articles. Each paragraph contains a specific kind of information (e.g., participant characteristics in a method's subsection called Participants). This also includes detailed guidelines for methodological and statistical reporting. For example, all dependent and independent variables should be disclosed and statistical analyses reported in detail (see Table III).

Disclosure of information for replications. The goal of open science is to make the complete research process transparent. Materials, data and code, and system software should be provided. This enables other researchers to replicate and reproduce results, but also to analyze data in a different way. The OSF provides the possibility to upload materials, data, code or any other files organized within projects for registered users. Authors can provide a link to the OSF project within the paper which can be blinded during the peer-review process. In addition, even if there was no preregistration or registered report, researchers can include the following sentence in their methods sections to label their research: "We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study." [18]. Open system software is an essential step for Affective Computing whose challenge lies in replicating studies with interactive systems. To reuse a system, program, or other research tools, other researchers should be able to re-build an interactive system. Such systems consist of several software components; some are freely available for research (e.g., software agent platforms, modeling standards), some of them are not (e.g., Text-to-Speech systems). Some systems rely on real-time hardware or software. Such systems might hinder replications — others might not have the required resources to obtain or use them. Finally, documentation on how to install and use such systems mainly does not exist for various reasons. Positive examples from other areas, however, do exist. For challenges of classifiers, such an open access practice is already implemented (e.g., *ACM Multimedia Grand Challenges*). Software environment containers, e.g., *Docker*, with documentation on how to start the software could facilitate the replication process. In Affective Computing, replicating studies from other laboratories is not a common standard so far. A prerequisite for replications is exactly this disclosure of materials and systems.

VII. Conclusion

This paper aims at providing an overview about the state of the art of empirical research in the Affective Computing community focusing on ACII'17 and '19 and at providing recommendations. By systematically analyzing the studies, we discovered that important information is missing for transparent, trustworthy, and replicable results. Those studies reporting

important statistical information, however, showed evidential value. Building on these results, we provide recommendations towards open research practices, envisioning improving the quality of studies in Affective Computing. We hope this analysis of research practices encourages the field of Affective Computing to move forward with regards to the important conversation of the credibility and trustworthiness of the community's research results.

REFERENCES

[1] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 981–995.

[2] S. E. Maxwell, M. Y. Lau, and G. S. Howard, "Is psychology suffering from a replication crisis? what does "failure to replicate" really mean?" *American Psychologist*, vol. 70, no. 6, p. 487, 2015.

[3] L. D. Nelson, J. Simmons, and U. Simonsohn, "Psychology's renaissance," *Annual Review of Psychology*, vol. 69, pp. 511–534, 2018.

[4] O. S. Collaboration *et al.*, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

[5] C. Allen and D. M. Mehler, "Open science challenges, benefits and tips in early career and beyond," *PLoS Biology*, vol. 17, no. 5, 2019.

[6] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein *et al.*, "The open science grid," in *Journal of Physics: Conference Series*, vol. 78, no. 1. IOP Publishing, 2007, p. 012057.

[7] D. B. Resnik, "Scientific research and the public trust," *Science and Engineering Ethics*, vol. 17, no. 3, pp. 399–409, 2011.

[8] C. Whitbeck, "Truth and trustworthiness in research," *Science and Engineering Ethics*, vol. 1, no. 4, pp. 403–416, 1995.

[9] B. Alberts, K. Shine *et al.*, "Scientists and the integrity of research," *Science*, vol. 266, no. 5191, p. 1660, 1994.

[10] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011.

[11] S. Stigler, "Fisher and the 5% level," *Chance*, vol. 21, no. 4, pp. 12–12, 2008.

[12] N. H. Steneck, "Fostering integrity in research: Definitions, current knowledge, and future directions," *Science and Engineering Ethics*, vol. 12, no. 1, pp. 53–74, 2006.

[13] J. Matthes, F. Marquart, B. Naderer, F. Arendt, D. Schmuck, and K. Adam, "Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013," *Communication Methods and Measures*, vol. 9, no. 4, pp. 193–207, 2015.

[14] N. Lynöe, L. Jacobsson, and E. Lundgren, "Fraud, misconduct or normal science in medical research–an empirical study of demarcation." *Journal of Medical Ethics*, vol. 25, no. 6, pp. 501–506, 1999.

[15] N. Butler, H. Delaney, and S. Spoelstra, "The gray zone: Questionable research practices in the business school," *Academy of Management Learning & Education*, vol. 16, no. 1, pp. 94–109, 2017.

[16] L. K. John, G. Loewenstein, and D. Prelec, "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological Science*, vol. 23, no. 5, pp. 524–532, 2012.

[17] K. Fiedler and N. Schwarz, "Questionable research practices revisited," *Social Psychological and Personality Science*, vol. 7, no. 1, pp. 45–52, 2016.

[18] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "Life after p-hacking," in *NA - Advances in Consumer Research*. Association for Consumer Research, 2013, pp. 17–19.

[19] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLOS Biology*, vol. 13, no. 3, pp. 1–15, 03 2015.

[20] G. Cumming, "The new statistics: Why and how," *Psychological Science*, vol. 25, no. 1, pp. 7–29, 2014.

[21] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, and M. A. L. M. van Assen, "Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking," *Frontiers in Psychology*, vol. 7, p. 1832, 2016.

[22] N. L. Kerr, "Harking: Hypothesizing after the results are known," *Personality and Social Psychology Review*, vol. 2, no. 3, pp. 196–217, 1998.

[23] M. Rubin, "When does harking hurt? identifying when different types of undisclosed post hoc hypothesizing harm scientific progress," *Review of General Psychology*, vol. 21, no. 4, pp. 308–320, 2017.

[24] ——, "The costs of harking," *The British Journal for the Philosophy of Science*, ahead of print.

[25] K. R. Murphy and H. Aguinis, "Harking: How badly can cherry-picking and question trolling produce bias in published results?" *Journal of Business and Psychology*, vol. 34, no. 1, pp. 1–17, 2019.

[26] B. A. Nosek and D. Lakens, "Registered reports: A method to increase the credibility of published results," *Social Psychology*, vol. 45, no. 3, pp. 137–141, 2014.

[27] A. E. van't Veer and R. Giner-Sorolla, "Pre-registration in social psychology — a discussion and suggested template," *Journal of Experimental Social Psychology*, vol. 67, pp. 2–12, 2016.

[28] M. Bakker, A. Van Dijk, and J. M. Wicherts, "The rules of the game called psychological science," *Perspectives on Psychological Science*, vol. 7, no. 6, pp. 543–554, 2012.

[29] M. Appelbaum, H. Cooper, R. B. Kline, E. Mayo-Wilson, A. M. Nezu, and S. M. Rao, "Journal article reporting standards for quantitative research in psychology: The apa publications and communications board task force report," *American Psychologist*, vol. 73, no. 1, p. 3, 2018.

[30] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement," *Journal of British Surgery*, vol. 102, no. 3, pp. 148–158, 2015.

[31] S. Mantha, M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss, "Comparing methods of clinical measurement: Reporting standards for bland and altman analysis," *Anesthesia & Analgesia*, vol. 90, no. 3, pp. 593–602, 2000.

[32] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin *et al.*, "Proposed minimum reporting standards for chemical analysis," *Metabolomics*, vol. 3, no. 3, pp. 211–221, 2007.

[33] *Publication Manual of the American Psychological Association: The Official Guide to APA Style*, 7th ed. American Psychological Association, 2020.

[34] E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan, "Promoting transparency in social science research," *Science*, vol. 343, no. 6166, pp. 30–31, 2014.

[35] K. Popper, *Conjectures and refutations: The Growth of Scientific Knowledge*. Routledge, 2014.

[36] N. C. Krämer, G. Lucas, L. Schmitt, and J. Gratch, "Social snacking with a virtual agent–on the interrelation of need to belong and effects of social responsiveness when interacting with artificial entities," *International Journal of Human-Computer Studies*, vol. 109, pp. 112–121, 2018.

[37] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, 2009.

[38] U. Simonsohn, L. D. Nelson, and J. P. Simmons, "P-curve: a key to the file-drawer." *Journal of Experimental Psychology: General*, vol. 143, no. 2, p. 534–547, 2014.

[39] U. Simonsohn, J. P. Simmons, and L. D. Nelson, "Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to ulrich and miller (2015)." *Journal of Experimental Psychology: General*, vol. 144, no. 6, p. 1146–1152, 2015.

[40] M. Brysbaert, "How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables." *Journal of Cognition*, 2019.

[41] C. Leys, M. Delacre, Y. L. Mora, D. Lakens, and C. Ley, "How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration," *International Review of Social Psychology*, vol. 32, no. 1, 2019.

[42] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.