Questionable Research Practices in Experimental Communication Research: A Systematic Analysis from 1980 to 2013

Abstract

Questionable research practices (QRPs) pose a major threat to any scientific discipline. This paper analyzes QRPs with a content analysis of more than three decades of published experimental research in four flagship communication journals: the *Journal of Communication*, *Communication Research*, *Journalism & Mass Communication Quarterly*, and *Media Psychology*. Findings reveal indications of small and insufficiently justified sample sizes, a lack of reported effect sizes, an indiscriminate removal of cases and items, an increasing inflation of *p*-values directly below $p < .05$, and a rising share of verified (as opposed to falsified) hypotheses. Implications for authors, reviewers, and editors are discussed.

Keywords: questionable research practices, experiments, effect size, confirmation bias

---

Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. *Communication Methods & Measures, 9*(4), 193-207.

Questionable Research Practices in Experimental Communication Research: A Systematic Analysis from 1980 to 2013

The 2011 article on false-positive findings in experimental psychology by Simmons, Nelson, and Simonsohn has set off an avalanche of discussions concerning so-called questionable research practices (QRPs), or "sloppy science". In the aftermath of this paper, QRPs have been examined in various fields and sub-disciplines, including medicine (Ionnadis, 2005; Jager & Leek, 2014; Ware & Munafò, 2014), psychology (Asendorpf et al., 2013; Bakker & Wicherts, 2014; Francis, 2014; Kühberger, Fritz, & Scherndl, 2014; Laws, 2013; Murayama, Pekrun, & Fiedler, 2014), criminology (Eisner, 2009), neuroscience (Chambers et al., 2014), education (Cook, 2014; Pigott et al., 2013), political science (Humphreys et al., 2013), management (O'Boyle et al., 2014) and consumer research (McQuarrie, 2014).

The concern most commonly expressed is that QRPs may result in a research bias: that is, "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (Ioannidis, 2005, p. 697). The definition and measurement of QRPs, however, is not straight-forward. Researchers may have reasons to engage in certain research practices, some practices may be justified in some contexts, and there certainly is a grey area (see also the Editorial by Vermeuleb & Hartmann in this issue). Nevertheless, selective or distorted reporting in experimental research can lead to a misinterpretation of results, with potentially damaging consequences for both the scientific community and society at large. The publication of inaccurate or misleading information may pose a threat because it can affect policy decisions, grant awarding, and ultimately decision making affecting individuals' lives (e.g. Wasserman, 2013).

Despite the immense relevance of this topic, no systematic analysis has been undertaken with regard to QRPs in experimental mass communication research. This is worrying, given that "QRPs are the steroids of scientific competition, artificially enhancing

performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage" (John, Loewenstein, & Prelec, 2012, p. 524).

The aim of this paper is to raise awareness of QRPs in communication research. More specifically, we report a systematic analysis of published experimental research based on a random sample drawn from four flagship communication journals: the *Journal of Communication*, *Communication Research*, *Journalism & Mass Communication Quarterly*, and *Media Psychology*. Our study covers three main areas: (1) sample size and statistical power, (2) case or variable deletion, and (3) confirmation versus falsification. These aspects are examined over a period of 33 years, thereby allowing us to observe changes over time.

Undisclosed Flexibilities in Experimental Research

Sample Size and Statistical Power

When conducting experimental studies, sample sizes are often reason for concern (e.g., Asendorpf et al., 2013; Ioannidis, 2005). In the context of QRPs, potential worries related to sample size include both small samples and the continued collection of data until results show up as hypothesized (i.e., significant; 'data-peeking', Funder et al., 2014). With small sample sizes, tiny changes in measures or models can lead to completely different results. Moreover, it has been argued that the "certainty (or precision) of an effect size estimate increases with greater sample sizes" (Lakens & Evers, 2014, 279), while smaller samples might reach larger, but less accurate, effect sizes (Kühberger et al., 2014). For that reason, larger samples are (generally) to be preferred (see also Murayama et al., 2014). The question of how sample sizes are employed and justified in published research is, therefore, of utmost importance.

Simmons et al. (2011) recommended that authors be precise about data collection termination rules *before* conducting a study and that they specify these rules in their papers. This may be achieved by, for instance, calculating *a priori* power analyses to determine the number of cases required for the hypothesized effect size (Bakker et al., 2012; Kühberger et al., 2014). However, "the rule is secondary, but it must be determined ex ante and be

reported" (Simmons et al., 2011, p. 1362). Independent of such a rule, as Simmons et al. (2011) argue, authors should collect at least 20 cases per cell in order to gain enough power for data analysis (e.g., 20 or more participants in each experimental group). If, for any reason, this is not possible, compelling justifications must be given for smaller cell sizes. For communication research, we have no systematic knowledge of the size of experimental samples or the justifications given for their selection.

Removal and Deletion

When it comes to justifying cell and group sizes, the removal of cases is especially problematic if no particular reason is given for their exclusion. While inconsistencies in reporting between N and group numbers should not raise suspicion in general, caution is definitely warranted: Bakker and Wicherts (2014), for example, found contradictions in the reporting of statistical values in 41% of their analyzed sample of psychological articles, with discrepancies appearing between reported sample sizes and degrees of freedoms (*df*) in *t* tests. The authors refer to this as a "contamination" (p. 6) of their meta-analysis, in so far as scholars omitted information about excluding cases from the data.

An extensive reporting of the design, measurements, and steps taken to clean and analyze data in general would help to improve the comprehensibility of results. Such steps would also facilitate the prevention of an outcome-reporting bias (Pigott et al., 2013): that is, the reporting of only those dependent measures that 'worked', after an initial loading of the experimental design with a large variety of variables. In a similar manner, Humphreys and colleagues (2013) refer to the reporting of solely the desirable (i.e., supported or significant) outcomes as 'fishing': a manner of noting only those results with the largest effect sizes, even when more than one possible outcome was initially tested.

In order to prevent such 'fishing', Simmons et al. (2011) urge scholars to report all variables measured in a study. The same applies to the conditions used in an experimental design: When manipulations fail or groups are merged during analysis, the changes must be

noted in the manuscript to prevent any selectivity in reporting 'good' outcomes only. Regardless of any reasons that may have led to the removal of outliers or other observations from the data, Simmons and colleagues (2011) urge all authors to report the statistical results for their analyses as if these eliminated cases were still included. While this approach may seem particularly restrictive, given the space that such additional results take up in the paper, it may add to the clarity and transparency of findings.

Confirmation versus Falsification

The practice of HARKing ("Hypothesizing After the Results are Known"; Kerr, 1998) has been a matter of debate for quite some time. HARKing refers to the habit of presenting post hoc hypotheses (i.e., developed after data analysis) as having been the scope of the study from the start. One 'advantage' of such a practice lies in the incentives connected to publishing results that are in line with a theory. However, as Kerr (1998) points out, the costs of HARKing for the development of scientific theory may be devastating – not least because it restrains researchers from "communicating valuable information about what did not work" (Kerr, 1998, p. 211).

As a consequence of HARKing and related practices, a positive research bias may emerge, insofar as significant (and positive) outcomes become more prominent in published work than non-findings or non-significant results (e.g., Bakker, van Dijk, & Wicherts, 2012; Francis, 2014; Ioannidis, 2005; Laws, 2013; Nelson et al., 2012). This bias has frequently been associated with the 'file-drawer-problem' of scientific work, in which 'undesirable' findings simply vanish, never to see the light of day. Francis (2014), for example, found that more than 80% of articles in *Psychological Science* presenting results of four or more individual experiments to be biased (i.e., inconsistent in terms of the probability of successful experimental outcomes). For psychological journals, Kühberger et al. (2014) detected an inflation of *p* values just below the significance level of .05, with significant findings being three times as frequent as insignificant ones (see also Masicampo & Lalande, 2012).

Given the ongoing debate surrounding the use and interpretation of *p* values, numerous authors have also stressed the importance of reporting accompanying effect sizes and 95% confidence intervals (e.g., Funder et al., 2014; Lakens & Evers, 2014; Leggett et al., 2013; Motulsky, 2014) and of not relying solely on significance levels when interpreting results.

Prevalence of QRPs

John et al. (2012) asked psychological researchers about their involvement in, as well as their acceptance (i.e., ethical defensibility) of QRPs (see also Fanelli, 2009). The results of their survey are worrying: Almost all (94%) of the respondents reported having engaged in at least one QRP, such as failing to report all of a study's dependent measures or deciding whether to continue data collection when results were insignificant. More than 25% of respondents admitted to having dropped a study's condition in reporting results, and even more respondents acknowledged having masked an unexpected finding as one predicted from the start (HARKing; Kerr, 1998; 27-35%, depending on the incentives for truth-telling). Moreover, when asked about the perceived defensibility of the practice, only one out of ten habits questioned in the survey (i.e., intentionally falsifying data) was perceived as unjustifiable. Furthermore, though doubts about one's own research integrity were relatively uncommon in the sample, more than 40% of psychologists said that they questioned the integrity of research conducted at other institutions at least "once or twice" or even "occasionally". Unarguably, as John et al. (2012) concede, QRPs cover a significant "gray area" of acceptable practices that blend into unjustifiable actions, and the line between solid and questionable practices may be difficult to draw (see also Fanelli, 2009).

In a survey by LeBel and colleagues (2013), the authors of articles in psychological journals submitted their rationales for omitting methodological information from their papers. Specifically, the QRPs questioned included the unreported exclusion of participants (e.g., due to missing data or outliers), the failure to disclose an expulsion of experimental groups (the

reason for which was stated as ‘unclear’ by almost 50% of respondents), the failure to report all measures tested (for reasons such as the measures being unrelated to the research question or non-significant), and the termination of data-collection (with almost 5% of researchers revealing that they collected data ‘until the pattern was clear’) (LeBel et al., 2013, p. 427). Along similar lines, Gordon (2014) explored the reasons for and possible execution of research misconduct by 581 full-time tenured and tenure track faculty from psychology and sociology departments at 40 U.S. research institutions. By presenting survey participants with different scenarios of QRPs (e.g., noncompliance with IRB, authorship considerations, false or adjusted reporting), she found fabrication and falsification of data to be rather rare among faculty members. However, ‘lighter’ forms of research misconduct were reported as being more likely to occur – and less problematic from a moral judgment perspective.

Comparable findings stem from a study conducted by Martinson and colleagues (Martinson, Anderson, & de Vries, 2005): Using survey data on early- and mid-career scientists, the authors note that “mundane ‘regular’ misbehaviours present greater threats to the scientific enterprise than those caused by high-profile misconduct cases such as fraud” (Martinson, Anderson, & de Vries, 2005, 737). Among these “regular” behaviors, the authors list the inadequate record keeping related to research projects as well as dropping observations or data points from analysis based on a “gut feeling”, using inadequate or inappropriate research designs, and the withholding of details of methodology or results. Finally, giving a more general overview, a meta-analysis of survey data on QRPs by Fanelli (2009) found that medical, pharmacological, and clinical researchers were significantly more likely to report on misconduct, as compared to biomedical researchers and scholars in other fields.

Surveys such as the ones reported above provide valuable insights into the prevalence of QRPs in different scientific disciplines and call to attention the fact that research misconduct covers a large area where differentiations between open fraud and cases of ‘minor adaption’ are very hard to make (e.g. John et al., 2012; Martinson et al., 2005). However, one

should not necessarily take the authenticity of self-reported behaviors and attitudes in this area for granted (Fanelli, 2009). Therefore, while surveys may allocate value to the detection and understanding of "sloppy science" practices, content analyses and meta-reviews are an important additional means for understanding prevalent research practices. Furthermore, surveys from other fields, such as psychology, cannot be generalized to the field of communication. We thus propose to systematically analyze the research published in major communication journals since the 1980s. Based on the current discussion in the field of psychology (e.g., Asendorpf et al., 2013; Bakker & Wicherts, 2014; Simmons et al., 2011) as well as other disciplines (e.g. Fanelli, 2009; Ionnadis, 2005) , we propose four general research questions:

RQ1: What are the sample sizes of reported studies, and how were the sample sizes determined and evaluated?

RQ2: To what extent were cases, outliers, or items removed?

RQ3: What is the ratio between verified and falsified hypotheses?

RQ4: Does the reporting of sample sizes, case, outlier, and item removal, as well as the ratio between verified and falsified hypotheses change over time?

Method

In a first step, we selected all articles (including research notes) that reported experimental studies from January 1980 to December 2013 in four major communication journals. We selected the *Journal of Communication* (N = 219 experimental papers; 257 single experiments), as the flagship journal of the ICA; *Communication Research* (N = 330 experimental papers; 397 experiments), as another top-tier journal with a strong focus on experimental research; and *Journalism & Mass Communication Quarterly,* as the leading AEJMC journal (N = 223 experimental papers; 240 experiments). We also included *Media Psychology* (N = 180 experimental papers; 243 experiments) as a key outlet for experimental research in the field. *Media Psychology* was analyzed from the first published issue in 1999.

For each journal, all articles that reported an experimental design study were selected. We did not follow the methodological interpretations of the authors; rather, we examined each research design to determine whether the used method met the standards of our definition of experimental design. We understand an experiment as a research design in which participants are randomly allocated to different conditions. Importantly, the treatment is manipulated by the researchers, allowing them full control over the processes of treatment construction and participant allocation. We also included quasi-experimental studies, which we understood to be studies in which researchers are not able to randomly allocate participants to different conditions.

In order to sample the experimental studies, the keywords, abstracts, and methods section of each article was read. In multi-study papers that reported more than one experimental study, each experiment was protocolled separately. This procedure resulted in a total of 952 articles reporting 1137 individual experiments.

Sample

For the purposes of the present paper, an analysis of all published experiments was not possible due to the time-consuming coding process. We drew a sample of n = 239 individual studies. Because the share of experimental studies was much lower in the 1980s than in the 1990s and the following years, we did not draw a random sample across all years. Instead, we randomly sampled approximately 70 experiments for each decade. Such a sampling strategy enabled us to compare different decades. In addition, 35 studies were randomly chosen from 2010 to 2013. This slight oversampling of the past four years allowed us a meaningful analysis of the present state of experimental research. Our final sample consisted of 46 individual experiments published in the *Journal of Communication*, 82 experiments in *Communication Research*, 74 experiments in *Journalism & Mass Communication Quarterly*, and 37 experiments in *Media Psychology*. Our main interest was to examine reporting practices and observe possible changes in these practices over time. However, our sample of n

= 239 individual studies was not suited to test interaction effects between journal and time (which was not our main interest). Therefore, we only report differences between journals for the full sample (from 1980 to 2013). We refrain from an arbitrary categorization of time into several periods for the statistical analysis, instead treating time as a continuous variable (MacCallum, Zhang, Preacher, & Rucker, 2002).

Measures and Coding Procedures

Based on the literature on QRPs discussed above, we developed a codebook with extensive variable definitions in several rounds of training involving five independent coders. We randomly selected experimental studies from the full sample in order to develop and refine our codes. After two extensive rounds of testing, during which categories were refined and some variables had to be dropped, we randomly selected ten experiments from the sample to determine inter-coder reliability. For most variables reported in this paper, reliability was satisfactory at this point. Two variables, however, could not reach acceptable levels of reliability (i.e., exact p-values shortly above .05 and effect size reporting). These variables were refined and tested in an additional reliability test using another random selection of ten experiments. This round of coding did not include those variables which had reached acceptable reliability levels in the first round. All articles used in the reliability tests were coded again in the final sample.

It is important to note that most variables involved extremely rare categories, for which the codes were almost always zero. In such instances, standard reliability coefficients are not informative due to the lack of variability. Therefore, in addition to Krippendorff's Alpha, we computed the chance-corrected reliability for detecting the most frequently coded category. Such a coefficient was suggested by Fretwurst (2015) and is called the standardized lotus. According to this method, when four coders coded zero and one coder coded one, we treated zero as the correct code due to it being the most frequently observed code. That is, the correct code serves as the reference category, or 'gold standard', based upon which the

chance-corrected reliability is computed. When there are only two coders, the Lotus is equivalent to Scott's Pi (see Fretwurst, 2015). For this study, we report the following variables from our codebook (see Appendix):

Samples. We coded whether authors conducted an *a priori* power analysis to determine the required "optimal" sample size before data collection (0 = no, 1 = yes; Lotus = 1.00, no variation). Then, we coded the size of the smallest experimental group (i.e., in five categories (1 = < 20, 2 = 20-29, 3 = 30-39, 4 = 40-49, 5 = ≥ 50; Alpha = .58; Lotus = .78), as well as whether at least one effect size was computed with respect to the formulated hypotheses (0 = no, 1 = yes for at least one hypothesis); Alpha = 1.0; Lotus = 1.00). Because samples and effect sizes may depend on the type of the experiment (i.e., within-subject or between-subject), we also coded the number of experimentally manipulated between (Alpha = .59; Lotus = .84) and within factors (Alpha = .72; Lotus = .89).

Case or Item Removal. We coded case deletion (0 = no, 1 = yes; Alpha = .59; Lotus = 1.00) and whether the results including the deleted cases where additionally reported in studies where cases were deleted (0 = no, 1 = yes; Lotus = 1.00, no variation). We also tracked the explicit mentioning of item deletion (0 = no, 1 = yes; Alpha = .73; Lotus = .94) and the availability of the data online (i.e., without the need to contact the authors; 0 = no; did not occur in the whole sample).

Confirmation versus Falsification. Finally, we counted the number of verified hypotheses, according to authors (including "partial support"; Alpha = .80; Lotus = .88), and the number of falsified hypotheses (Alpha = .72; Lotus = .87). We also coded the occurrence of any exact *p* values of just below .05 (i.e., .040 - .049; Alpha = .73; Lotus = .97) or just above .05 (i.e., .050 - .059; Lotus = 1.00) and whether at least one confidence interval (CI) was reported (Alpha = 1; Lotus = 1.00).

Results

We calculated the percentage of articles reporting experiments in comparison to the percentage of articles reporting other methods. As presented in Figure 1, experiments are gaining increasing importance in all four journals. From 1980 to 1984, only 7.52% (n=24) of all articles published in the *Journal of Communication* were experimental studies. By 2010 to 2013, the share had increased to 49.28% (n=103). In *Communication Research*, the share of published experimental studies rose from 14.63% (n=18) from 1980 to 1984 to 53.85% (n=77) from 2010 to 2013. Likewise, only 7.42% (n=38) of all published articles in *Journalism & Mass Communication Quarterly* reported an experimental approach from 1980 to 1984. In more recent years (2010 to 2013), 19.89% (n=29) of all published articles have been experiments. In this instance the number of overall articles decreased from 512 in the period of 1980 to 1984 to 131 in the period from 2010 to 2013. In *Media Psychology*, 42.11% (n=8) of all published articles reported experiments in 1999. In 2013, the share of experiments in the same journal rose to 80% (n=16).

Samples and Effect Sizes

With regard to the number of participants, only 4 of the 239 experimental studies conducted *a priori* power analyses in order to determine sample sizes. A considerably large share of studies had sample sizes with fewer than 20 participants (25.5%) in the smallest experimental group, and 22.2% employed samples with 20-29 subjects. In 10.9% of all experiments, the samples included 30-39 subjects; in another 10.9%, the number was between 40 and 49; and 30.1% of all sampled studies involved more than 50 participants in their smallest samples. The size of the smallest experimental group did not significantly differ by journal (chi-square test, $\chi^2 = 18.73$, $df = 12$, $\Phi = .28$, $p = .10$). However, sample size in the experimental groups decreases with the number of within- (ordinal logistic regression, $b = -.54$, Wald = 12.42, $p < .001$) and between-factors ($b = -.97$, Wald = 34.19, $p < .001$; Nagelkerke $R^2 = .17$).

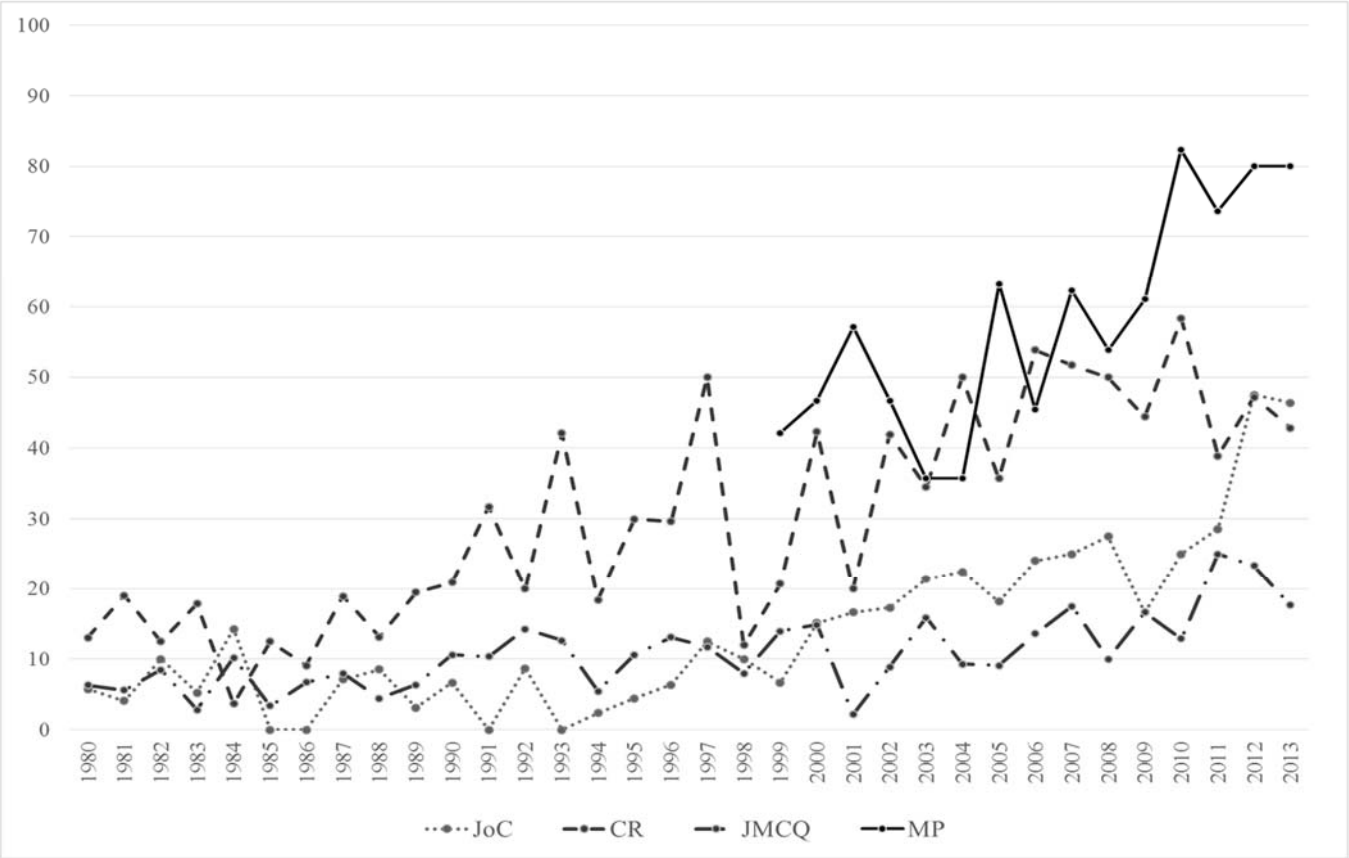

FIGURE 1. Share (%) of articles reporting an experiment from 1980–2013.

Note: JoC = Journal of Communication, CR = Communication Research, JMCQ = Journalism & Mass Communication Quarterly, MP = Media Psychology.

To examine whether sample sizes changed over time, we conducted an ordinal logit regression analysis using the size of the smallest cell as the dependent variable and the year of publication (1 = 1980; 34 = 2013) as the independent variable. We found no relationship between sample size and time (b = .01, Wald = 1.24, p = .27; Nagelkerke R^2 = .01). We did not calculate interaction effects between time and journal due to the small sample sizes of each journal for individual years.

A considerably large share of studies reported effect sizes (57.3%) in relation to hypotheses. As a binary logistic regression revealed, the tendency to report effect sizes increased significantly over time (b = .09; Wald = 34.97; p < .001; Nagelkerke R^2 = .21). For instance, when looking at 2010-2013, 94% of the studies reported effect sizes (compared to

22% from 1980-1983). Reporting of effect sizes did not depend on the number of within- (b = -.05; Wald = .13; p = .72) and between-factors (b = .16; Wald = .12; p = .29; Nagelkerke R^2 = .01). Moreover, the reporting of effect sizes differed across journals: Whereas 73.2% of all experiments published in *Communication Research*, 67.4% in the *Journal of Communication*, and 64.9% in *Media Psychology* reported effect sizes, only 29.7% did so in *Journalism & Mass Communication Quarterly.* The difference between the latter journal and the other three was highly statistically significant (chi-square test; $\chi^2 = 34.22$, $df = 3$, $\Phi = .38$, $p < .001$).

Case or Item Removal

In 22.2% of all studies (n = 53), the authors admitted to having excluded cases. Among those, n = 5 stated that the excluded subjects identified the purpose of the study and were therefore removed, n = 4 noted unlikely response patterns, n = 3 referred to outliers, n = 11 gave multiple reasons, and n = 26 mentioned other specific reasons (which were not coded in detail). In only four cases, no reason was given for case removal. Interestingly, the likelihood of reporting case removal increased over time (binary logistic regression model; b = .03; Wald = 3.93; $p < .05$; Nagelkerke R^2 = .03). Among those studies that excluded cases, only n = 2 experimental studies provided results that included the deleted cases. Similarly, only 5% of all analyzed studies reported findings that also included the deleted items.

Confirmation versus Falsification

As could be expected, the average number of verified hypotheses (M = 2.26, SD = 2.19; ranging from 0 to 14) was much higher than the number of falsified hypotheses (M = 0.77, SD = 1.15; ranging from 0 to 6). Interestingly, the ratio between the number of verified and falsified hypotheses grew more positive over time (ordinal logit regression; b = .06; Wald = 22.78, $p < 001$; Nagelkerke R^2 = .09). This indicates that, over time, more hypotheses were verified than falsified. The ratio was significantly ($p < .001$) more positive in *Communication Research* (M = 1.93, SD = 2.68) than in *Journalism & Mass Communication Quarterly* (M =

.89, *SD* = 1.95). No differences were observed for the other two journals (*Journal of Communication*: *M* = 1.78, *SD* = 2.42; *Media Psychology*: *M* = 1.32, *SD* = 2.19).

Finally, we looked at the frequency with which communication research scholars reported *p*-values just below or above *p* = .05. We observed that only 11.7% of all studies reported *p*-values just below *p* = .05. Roughly the same share of studies reported *p*-values just above *p* = .05. When looking at how the ratio of just below to just above changed over time, we observed that, over time, p-values were more likely to be just below than just above p = .05 (ordinal logit regression; b = .05, Wald = 5.56, $p < .05$; Nagelkerke $R^2 = .04$; CI from .01-.09). Only three of the N = 239 studies reported confidence intervals.

Discussion

The world of academia has witnessed an increasing awareness of QRPs, such as problematic sample sizes, low statistical power, undisclosed flexibilities in the selection of items and cases, and confirmation biases in published research (e.g., Asendorpf et al., 2013; Bakker & Wicherts, 2014; Chambers et al., 2014; Humphreys et al., 2013; Jager & Leek, 2014; O'Boyle et al., 2014; Pigott et al., 2013). In some disciplines, such as psychology, major associations have launched landmark initiatives to improve the quality and transparency of scientific research (e.g., Funder et al., 2014). Yet, in the field of experimental communication research, neither journals nor academic associations have followed this trend. Considering the fact that more than one third of all studies published in the field's flagship journals, such as the *Journal of Communication,* report experimental designs, a reflection and description of potentially questionable research practices is long overdue.

Our findings suggest that, on the one hand, a large share of experimental studies still work with considerably small sample sizes. More importantly, this has not changed since the 1980s. In addition, more than one third of all studies in our sample did not report effect sizes. Although the reporting of effect sizes has increased over the years, this finding is still

alarming. On the other hand, while slightly increasing over time, unjustified case removal is still the exception in experimental communication research.

Finally, it was not surprising to find that studies were more likely to report verified than falsified hypotheses, although it should be noted that the ratio between verification and falsification has become more positive over the years. Similarly, the ratio between reported *p*-values just below and just above $p = .05$ has become more positive, indicating an increasing inflation in *p*-values around the infamous .05 mark. When looking at the whole period from 1980 to 2013, the overall number of *p*-values just below $p = .05$ was roughly equal to the number of *p*-values just above. However, when only taking into account those studies published after the year 2006 (n = 58), which is roughly comparable to the sample of Kühberger and colleagues (2014), the pattern is different: 17.2% of our sampled studies reported *p*-values shortly below $p = .05$ while only 5.2% reported *p*-values shortly above $p = .05$. Thus, even though our considerably small sample size needs to be taken into account, our findings equal those obtained by Kühberger et al. (2014) who observed—for psychological articles from all areas of psychological research published in the year of 2007—"about 3 times as many studies just reaching than just failing to reach significance" (p. 5).

Our results have three different sets of implications for authors, journal editors, and reviewers, all of which have already been proposed elsewhere (Bakker et al., 2012; Chambers et al., 2014; Funder et al., 2014; Kühberger et al., 2014; Simmons et al., 2011). First, when it comes to study design and implementation, we echo the call made by many others to conduct *a priori* power analyses to determine the designated number of cases required for the hypothesized effect sizes (see Bakker et al., 2012; Kühberger et al., 2014; Simmons et al., 2011; Thorson, Wicks, & Leshner, 2012). Moreover, sample sizes below 20 cases should be regarded with caution, and all observed effect sizes should be reported.

Second, more sensitivity is in order with regard to case or item deletion. Clear reasons for such actions should be provided, and, whenever possible, scholars should report findings

that include the eliminated cases. Needless to say, practices such as dropping observations or measures until findings turn out as expected and running studies until effects become significant are highly problematic (Simmons et al., 2011).

Third, and related to the above, all procedures, whether intentional or unintentional, that favor verification over falsification in conducting and publishing research may undermine our ability to draw valid conclusions about communication phenomena, distort meta-analytic reviews, and can lead to misguided theoretical debates and research directions (see Funder et al., 2014). As has been stated numerous times in other disciplines, we need to give null findings and strict replications a valued home. We encourage journal editors to establish so called "replication corners" and to instruct their reviewers about the value and necessity of replications. One key necessity is that both data and experimental material should be openly available to the scientific community for the purpose of facilitating replication (e.g., Bakker & Wicherts, 2014; Chambers et al., 2014; Funder et al., 2014; O'Boyle et al., 2014). Moreover, reviewers and journal editors should be more tolerant in the face of "imperfections" (i.e., insignificant results or non-findings), which may well add to the state of knowledge – even though (or exactly because) they were not perfectly crafted in theory (Simmons et al., 2011).

In another possible strategy, some journals (e.g., *AIMS Neuroscience; Attention, Perception & Psychophysic; Cortex*; *Experimental Psychology*; *Social Psychology*) have implemented a policy that requires authors to pre-register their hypotheses before data collection "in order to more credibly distinguish hypothesis testing from hypothesis generation" (Miguel et al., 2014, 30; see also Asendorpf et al., 2013; Chambers et al., 2014, Humphreys et al., 2013; Ioannidis, 2005; Leggett et al., 2013). In a two-stage review process, authors first submit for peer review a full study proposal that includes all hypotheses. Following acceptance of this protocol, the authors conduct the study and then resubmit the paper, including the results of the pre-registered analyses and other exploratory analyses, which are clearly separated from the pre-registered ones (see Chambers et al., 2014). The idea

is that exploratory research should be given the status it deserves, but not be 'sold' as having been theorized from the very start. In other words, "It is an illusion to believe that we gain something of value by insisting that authors pretend to have 'known all along'" (Kerr, 1998, p. 216; see also Wagenmakers et al., 2012).

There are a number of limitations to our study that are worthy of careful consideration. To begin with, some variables that were originally included in our codebook failed to achieve sufficient levels of intercoder reliability. For instance, we were unable to code whether a study was designed as a strict or conceptual replication. It was also impossible to determine the inclusion of covariates in statistical models in a reliable way. Coding experimental studies requires researchers to fully read each paper, which is an enormous effort. Some information is hidden in footnotes; other information is only implied or stated in sections of the paper where it does not belong. These difficulties did also affect our coding of case removal. More specifically, it was not possible to determine the precise number of excluded cases because these could vary within a study. We faced similar problems in determining the exact sample size of (the smallest) experimental groups. In a similar manner, the reasons given by researchers for case removal should be analyzed in more detail than it has been done here.

Moreover, for most codes in our codebook, we did not follow the judgment of the authors; instead, we used our own definitions detailed above. On the one hand, this is the only way to arrive at valid conclusions. On the other hand, this makes coding extremely labor-extensive. Even more importantly, some QRPs – such as, for instance, running an experiment multiple times until a desired outcome is found – may have occurred but not been reported. In line with this argument, caution is warranted when it comes to the ability of content analyses to detect QRPs. Our coding of articles had to rely on the information provided by the authors. Accordingly, some scholars have turned to survey measurements instead, asking researchers about their use of such practices (e.g. Fanelli, 2009; Gordon, 2014; John et al., 2012; LeBel et al., 2013; Martinson et al., 2005). Albeit possibly biased in terms of self-reported behavior,

we strongly encourage future studies in our field to employ survey measures to analyze the prevalence of QRPs as well as their acceptance and the reasons for engaging in them.

In sum, our content analytical approach cannot prove QRPs in the published communication literature. However, this was by no means our aim. Also, we do not want to question the integrity and validity of experimental communication research in general or of single studies. In fact, our findings suggest that there are some indications for QRPs in our discipline. However, we are very far from concluding that QRPs are prevailing our field. Our aim was to raise awareness for QRPs in experimental communication research. It is our hope that this study will spur discussion among authors, reviewers, and editors, potentially leading the way to improved and more transparent research practices.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in Psychology. *European Journal of Personality, 27*(2), 108-119.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE, 9*(7), e103360.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Editorial: Instead of "playing the game" it is time to change the rules: Registered reports at AIMS Neuroscience and beyond. *AIMS Neuroscience, 1*(1), 4-17.

Cook, B. G. (2014). A call for examining replication and bias in Special Education Research. *Remedial and Special Education, 35*(4), 233-246.

Eisner, M. (2009). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology, 5*(2), 163-183.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5). doi: 10.1371/journal.pone.0005738

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*(5), 1180-1187.

Fretwurst, B. (2015). Reliabilität und Validität von Inhaltsanalysen. Mit Erläuterungen zur Berechnung des Reliabilitätskoeffizienten ›Lotus‹ mit SPSS. In W. Wirth, K. Sommer, M. Wettstein, & J. Matthes (Eds.), *Qualität in der Inhaltsanalyse* (pp. 176-203). Cologne: Halem.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in Personality and Social

Psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*(1), 3-12.

Gordon, A. M. (2013). Rational choice and moral decision making in research. *Ethics & Behavior, 24*(3), 175-194.

Humphreys, M., Sanchez de la Sierra, R., & van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis, 21*(1), 1-20.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics, 15*(1), 1-12.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in Psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE, 9*(9), e105825.

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278-292.

Laws, K. (2013). Negativland - a home for all findings in psychology. *BMC Psychology, 1*(1), 2.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org : Grassroots support for reforming reporting standards in Psychology. *Perspectives on Psychological Science, 8*(4), 424-432.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology, 66*(12), 2303-2309.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737-738.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

McQuarrie, E. F. (2014). Threats to the scientific status of experimental consumer psychology: A Darwinian perspective. *Marketing Theory*. Advanced online publication.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Van der Laan, M. (2014). Promoting transparency in Social Science Research. *Science, 343*(6166), 30-31.

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology, 387*(11), 1017-1023.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*(2), 107-118.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry, 23*(3), 291-293.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2014). The Chrysalis Effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. Advanced online publication.

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in Education Research. *Educational Researcher, 42*(8), 424-432.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive Psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Thorson, E., Wicks, R., & Leshner, G. (2012). Experimental methodology in Journalism and Mass Communication Research. *Journalism & Mass Communication Quarterly, 89*(1), 112-124.

Vermeulen, I. E., & Hartmann, T. (2015). Questionable research and publication practices in communication science. *Communication Methods & Measures, 9*(4), 189-192.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.

Ware, J. J., & Munafò, M. R. (2014). Significance chasing in research practice: Causes, consequences and possible solutions. *Addiction, 110*(1), 4-8.

Wasserman, R. (2013). Ethical Issues and guidelines for conducting data analysis in psychological research. *Ethics & Behavior, 23*(1), 3-15.

Appendix

Codebook

1.	A priori power analysis; authors describe having determined required "optimal" sample size before data collection based upon power analysis	0 no 1 yes
2.	Sample size of groups; give size of smallest experimental group (calculate [total sample size/number of groups] if not reported), even before groups are merged (e.g. when question order is a between-subjects factor, but shows no effects)	1 < 20 2 20-29 3 30-39 4 40-49 5 >/= 50
3.	Effect sizes reported concerning hypotheses? (Pearson's r, Cohen's d, Eta square, R square, Cramer's V, explained variance, squared multiple correlation, Epsilon squared; check tables as well, not only text) Statistical abbreviations: r; d; η^2; R2; V or ϕ or φ [Phi]	0 no 1 yes, for at least one hypothesis
4.	Number of experimentally manipulated between subjects factors (not age, sex, education etc., i.e. factors that cannot be manipulated)	0-
5.	Number of experimentally manipulated within subjects factors (i.e. the same participant is exposed to different stimuli; not age, sex, education etc., i.e. factors that cannot be manipulated))	0-

6.	Cases excluded/not analyzed on purpose (according to the authors) Not: Missing values	0 no 1 yes
7.	Reason given for case deletion	1 unlikely response patterns 2 participant discovered purpose of the study 3 outliers 4 multiple 5 other (mentioned) 6 other (not mentioned, i.e. unclear) *77 not applicable (no cases deleted)*
8.	If cases were deleted, results given including deleted cases?	0 no or not applicable (no deleted cases) 1 yes
9.	Items deleted?	0 no 1 yes (at least one)
10.	Data: Dataset/scripts available online (without the need to contact the authors; i.e. authors provide a link for the source of the data within the text or cite public data such as Eurobarometer etc.)	0 no 1 yes
11.	Any exact p-values reported shortly below .05 (.040 - .049)? Do not consider possible round ups.	0 no 1 yes

12.	Any exact p-values reported shortly above .05 (.050 - .059)? Do not consider possible round ups.	0 no 1 yes
13.	At least one confidence interval (CI) reported?	0 no 1 yes