Using an inquiry into data integrity to improve future research in behavioral science

******* PRELIMINARY DRAFT - PLEASE, PLEASE DO NOT CITE OR CIRCULATE *******

Michael Yeomans[1*], Karen Huang[2], Martha Jeong[3], Hanne Collins[4],
Julia Minson[5], Alison Wood Brooks[5], and Mark Kennedy[1]

[1]Imperial College London
[2]Georgetown University
[3]Hong Kong University of Science and Technology
[4]University of California - Los Angeles
[5]Harvard University

* Corresponding Author:
Michael Yeomans
m.yeomans@imperial.ac.uk

Abstract

In recent years, many questions have been asked about the credibility of behavioral science research. However, the response from the field has been scattered and small. Verifying the scientific record, especially for research conducted long ago, is difficult. We argue for a more forward-thinking approach. For too long, data integrity in our field has been taken for granted by producers and consumers of research. This leads researchers to routinely overlook simple efforts that could improve the credibility of their work. As a case study, we conduct an audit of three recently published papers. We cooperated with a team of independent data scientists to verify every step of our workflow, from raw data collection to final published results. This process confirmed all of the substantive findings of this research–but importantly, our audit also revealed several avoidable vulnerabilities in our research workflow: gaps in record-keeping, over-reliance on sensitive data, oversights in analysis code, weaknesses in common software tools, and difficulties replicating software environments. We offer recommendations for individuals designing and conducting their research, for auditors investigating past work, and for institutions setting policies to encourage more careful consideration of data integrity in the future. Although fraud and error cannot be eliminated entirely, researchers and institutions can greatly reduce the risk of both by proactively incorporating integrity concerns into their workflow in advance.

Keywords: Open science, Reproducibility, Science of science

Introduction

Scientific misconduct is immeasurably costly. False findings mislead future researchers into pursuing follow-up research that is hopeless. They also mislead practitioners and lay audiences to change their beliefs and behavior in erroneous ways. There are opportunity costs, as resources and attention are drawn away from credible research. And there are reputational costs. When trust is undermined, it can undermine the credibility not just of the people who committed misconduct, but also their communities, their institutions, and even their entire field of inquiry.

At the same time, investigating scientific misconduct is also costly. It can be almost as hard for outsiders to verify a result as it is to generate it in the first place. Further, accusations of scientific misconduct are taxing for the accused, the investigators, and their colleagues. It is no wonder these investigations are rarely undertaken. Instead, resource allocation decisions (e.g. journal pages, grant funds, awards, hiring) depend heavily on subjective assessments of trust. In recent years, these levels of trust afforded have come into question.

In this paper, our aim is to present a clear set of steps that authors can take to improve the credibility of research conducted going forward. We do this by presenting an audit we conducted to verifiably establish trust in three published papers. This audit was initiated when a set of research misconduct allegations came to light in the field of behavioral science in June 2023 (though this has not been the only case of its kind, e.g. Stroebe, Postmes & Spears, 2012).We offer no comment on the validity of the allegations, the ensuing conflicts and legal actions, or the credibility of any of the research in question. Rather, our hope with this paper is to contribute to a future of greater research integrity and credibility in behavioral science.

Misconduct cases typically provoke responses that are mostly backward-looking– focusing on which past papers may have been fraudulent and what are the appropriate sanctions. And while the responses to such cases are initially focused narrowly on a specific set

of research papers in question, the consequences are much broader. These cases raise suspicions and doubts about other research as well–the whole body of work for the individual in question, other papers authored by their colleagues (Mongeon & Lariviere, 2016), and even their field in general. We believe much of this broader reaction is not based on evidence and thus is likely unwarranted, even if it is representative of an understandable desire for trustworthy research. Still, such doubts are hard to quell, even for honest researchers, because there are no objective or established protocols for affirming the integrity of behavioral research.

In the wake of the 2023 case, relatively little has been said about how to proceed. Here, we offer a forward-looking approach. Anecdotally, we believe behavioral science workflows routinely overlook practical measures for prioritizing data integrity, and this raises the risk of inaccurate results either through honest errors, deliberate misconduct, or gray areas in between. To date, virtually all assessments of integrity are based on self-report, which is an unreliable measure of unethical behavior (John, Loewenstein & Prelec, 2012). Our audit workflow systematically addresses these weak points. We hope to improve not only assessments of past research, but also the way in which future research is conducted, documented, evaluated, and disseminated (see Figure 1 below).

Methods

In the last 10-15 years, there have been many positive developments toward improving the credibility of the scientific body of knowledge, targeting various stages in the research production pipeline. Our audit is complementary to, but does not address, concerns such as replicability (can a result be demonstrated in new data from the same population; Nosek et al, 2022), generalizability (can a result be demonstrated in new data from a different population; Simons, Shoda & Lindsay, 2017; Yarkoni, 2022), and measurement validity (do the collected data represent the underlying construct; Flake & Fried, 2020). Instead, our audit framework focuses on two core components of the research production pipeline:

- reproducibility - can a result be verified using the provided data?
- integrity - do the provided data accurately reflect the data generating process?

Reproducibility audits take the provided data at face value, and evaluate whether the published results can be generated by an outside researcher using that data (Natl. Acad. Sci. Eng. Med., 2019). Ideally this is done using analysis code provided by the authors, though in principle a sufficiently detailed description of the analysis could suffice. Considerable efforts have been made to encourage and facilitate the public posting of data and analysis code by researchers, for example the Open Science Framework ("OSF", Munafo et al., 2017). Journals are increasingly requiring this as a condition for publication (Simchi-Levi, 2023; Hardwicke & Vazire, 2024). Additionally, there have been recent efforts to systematically audit the reproducibility of many papers at once (e.g. Artner et al., 2021; Nust & Eglen, 2021; Crüwell et al., 2023; Fisar et al., 2024).

These are welcome developments–in fact, many of the open research practices that have taken hold over the last decade are complementary to the questions at the heart of this audit. Reproducibility is necessary for credibility, but we believe it is not sufficient. By design, reproducibility audits assume the provenance and integrity of the data. However, research misconduct cases call into question the authenticity of *the data itself*. Thus, a deeper evaluation of the integrity of the workflow is required. Fraudulent datasets can reproduce fraudulent results–but this does not speak to whether the data provided reflect the true data generating process.

In this paper, we therefore introduce a procedure for conducting an "integrity audit": practical routines for evaluating how faithful posted data are to the actual data-generating processes. Problems of data integrity can arise from a variety of unintentional errors during data collection and cleaning (e.g. Herndon, Ash & Pollen, 2014; Yeomans et al., 2019). Even worse, data integrity is threatened when researchers distort their results by systematically altering data to support their desired conclusions. Such distortions generally involve data tampering practices

such as arbitrarily excluding participants, manually editing measurements, or wholly fabricating data. Researchers regard data tampering as categorically worse than analysis-altering behavior (John, Loewenstein & Prelec, 2012). However, data integrity has received relatively little attention compared to the growing literature on "p-hacking."

Sample

Our audit focused on three published papers co-authored by Dr. Francesca Gino and the lead author of this paper: "It Doesn't Hurt to Ask: Question-Asking Increases Liking" ("QA-2017"; Huang, Yeomans, Brooks, Minson & Gino, 2017); "Communicating Warmth in Distributive Negotiations is Surprisingly Counter-Productive" ("CW-2019"; Jeong, Minson, Yeomans & Gino, 2019); and "Conversational receptiveness: Improving engagement with opposing views" ("CR-2020", Yeomans, Minson, Collins, Chen & Gino, 2020). The three papers include 16 separate empirical studies, which are listed in Table 1.

These three papers are not a representative sample of behavioral science. However, they form an excellent case study for our workflow, for two reasons. First, their design makes auditing possible. By possible, we mean that almost all the data was collected via a digital platform (e.g. Qualtrics). Thus the raw data was recorded and stored externally, which is not true in many behavioral studies (e.g. paper and pencil surveys). Furthermore, with rare exceptions, the analyses and data cleaning were conducted in the open-source R programming language. This leaves a record of exactly how the data were handled.

Although these papers are possible to audit, they are not easy to audit, and thus we think they represent a challenging "stress-test"' of the proposed audit process. First, each paper studied interactions between multiple people (e.g. negotiations, conversations), which complicates recruitment, data linkage, and exclusion criteria. Second, multiple digital platforms (e.g. Qualtrics, Amazon Mechanical Turk, Chatplat, iDecisionGames) were used within each publication, necessitating multiple approaches to grant full access to the auditors. Third, much

of the resulting data was unstructured (i.e. natural language), and in some cases had to be manually processed by research assistants (e.g. spell-checking). Finally, in some cases, these data were collected many years ago. As time passes, authors often change institutions, with insufficient systems in place to transfer or access data.

Protocol

We divided the auditing work for each paper in two steps. Each paper was evaluated separately for *reproducibility* (can the results in the paper be derived from the clean data posted on OSF?) and *integrity* (can the clean data be derived from the raw data source?), for a total of six audit reports (2 reports x 3 papers). Each report consisted of a checklist, completed by an external auditor, which was then reviewed by the authors of the original paper (to clarify questions and revise the code base), and then finally by senior external supervisors.

Reproducibility. The reproducibility audits were inspired by existing frameworks from the literature (primarily Fisar et al., 2024). For the current papers, because all of the data and analysis code for every study in each paper were already posted to OSF, the auditors merely had to download the data and code and conduct the same analyses, line by line, and evaluate whether the produced results matched what was reported in the paper. The result is a checklist (see Table 2), which can be applied separately to each study in a paper.

The studies across these three papers were not perfectly reproducible. There were several cases where results reported in the paper were missing from the analysis code, or mismatched with the code output. As a result, each external reproducibility audit was completed twice - the original authors used feedback from the first audit to revise the code base and resubmit their work to the auditors a second time. This revision process was modeled on the original design of the Fisar et al. audit (though they did not end up completing the second round due to lack of resources). We recommend this collaboration be a part of the audit process

because by permitting revisions, this approach to auditing is more forward-looking, and prioritizes the quality of the cumulative scientific record rather than defense of past practices.

Integrity. Due to the paucity of existing work on data integrity, we developed a novel framework for these audits. We again established a checklist, isolating each step in the workflow to produce the cleaned datasets from the raw data (see Table 2). In practice, all cleaning code could be posted to OSF, mirroring the reproducibility audit. However, access to raw data cannot be shared so easily. First, raw data can be altered before being posted. Furthermore, some raw data should never be posted, because it contains personally-identifying information ("PII"). We posted PII-cleaned data to OSF, for posterity.

For the integrity audit, we arranged full access to the raw data for the auditors. Each data source required a different approach, which we describe in detail in Appendix A. For archival data borrowed from other authors, we passed along the shared data, which at least confirms that we did not mishandle it ourselves. For data collected on digital platforms (Qualtrics, iDecisionGames, ChatPlat, Google Workspace) we provided login information to the auditors, to see the original version of the data stored online, along with records of any post-collection changes to the data. Only one dataset eluded this level of scrutiny.[1] Where we relied on human data entry (annotating text, applying exclusion criteria), we provided the original Excel files. However, this does not fully exclude the possibility of data tampering. By design, these tasks were performed blinded to information that could bias results (outcomes, conditions), but

[1] In Study 2 of QA-2017, Mechanical Turk ID logs were used in the cleaning code to filter out participants who joined the study multiple times. After many months of requests, we finally convinced the school to briefly re-activate her old email account so she can transfer the Mechanical Turk account to her new email address. We are currently trying to reactivate her old Mechanical Turk account but there is no guarantee the data will still be stored, even if it is reactivated. Instead, the auditors had to rely on the copies of the Mechanical Turk worker logs (stored on Dropbox since 2015).

we cannot guarantee this information did not leak in other ways. In these cases, trust in our double-blind protocol is required.

Results

The full text of the six reports are shared online at https://osf.io/rymv8/. Overall, the audit found no threats to the underlying integrity of any of the data in the three papers, and all of our substantive conclusions remain valid. There were cases where the results deviated slightly between the original paper and the corrected analyses. However, all estimates of statistical significance and effect sizes were not meaningfully different.

For the authors of these papers, our audit gives us the assurance to confidently stand by all of the conclusions of the three papers, in total. We encourage other researchers to continue to cite our work, teach it in classes, and re-use the data and code in their own research. With this audit, these papers have undergone an unusual level of scrutiny that may make them more credible than similar papers that have not been audited (regardless of authorship).

That said, this audit also revealed aspects of our workflow that were not ideal, especially by current norms. As behavioral research has been developing and adopting new standards for integrity and reproducibility—such as pre-registration and other open science practices—over the last decade, our own practices have also changed, and this is reflected in the three papers considered here. Compared to the first two papers, CR-2020 relied less on human data cleaning and had far fewer gaps in the analysis code (it also included several pre-registered studies). As a result, the current analysis code in CR-2020 more closely matches the paper that was published (although there are still some deviations). We believe the improvement in research hygiene across these three papers tracks the positive developments in behavioral research methods in the field over the last decade.

Isolated Errors. Some deviations seem to have arisen from honest error that could be characterized as lapses of carefulness. For example, in two places in QA-2017, a Cronbach's

Alpha was described as an intraclass correlation. There were several places where a number was off by a significant digit, and sometimes two (perhaps typos). A negative sign was left off an interaction term. One incomplete observation from Study 1A of CR-2020 was not part of our download of the raw data, and was thus not counted among our reported exclusions. Most troublingly, we could not find any code for Table 3 from QA-2017, and the numbers we reproduced during the audit were all different, apart from the sample size. The substance of the conclusions remain unchanged (in fact, the statistical evidence was often stronger) but these were avoidable errors. To rectify these errors, we have submitted a correction to QA-2017 to the journal, so that the official version of the paper matches what the code currently produces. There were fewer of these issues in CW-2019, and almost none in CR-2020. This reflects our eager adoption of field-level changes to our workflow. As analysis code became a more focal part of scientific contributions, we responded by having multiple authors check the code before submission.

Systematic Errors. Other concerns cropped up repeatedly across papers, and these indicate clear room for improvement in our future research. For example, we did not write rounding into the R code, instead rounding by hand. This perhaps contributed to transcription errors where the last significant digit in the paper was incorrect. We also never specified random seeds in our original scripts. This meant any analysis dependent on randomness (e.g. cross-validation, bootstrapping) was not exactly reproducible. For the audit, we reverse-engineered most of the original random seeds after the fact, by testing thousands of random seeds until one produced the exact original results - but this is *not* recommended. Always set the random seeds in your code.

We also had difficulty recreating the original software environments. Primarily, CW-2019 and CR-2020 use the politeness R package (Yeomans, Kantor & Tingley, 2018), which was under continuous development as the papers were being written. Older published versions of the package could not generate every exact number from the paper, suggesting some

intermediate, non-published versions were used for parts of each paper. Instead, we chose an older version that was published near in time to the final writing of the manuscript to conduct the audit. To be clear, the correlations between the measures used for the paper and the measures reproduced in the audit were always above 0.95. And the substance of all results held, even with the newest version of the package. Furthermore, the package has improved over time, so it is not in fact clear which set of results is most accurate. Still, this meant almost all of the NLP results in the audit reproduction were slightly different than the ones in the original paper. This reduced the overall reproducibility of the paper, and complicated the auditors' work. Instead, we recommend authors only use stable, published versions of packages for the final manuscript, and make those versions (and the steps for installing them) explicit in their code.

Audit Limitations. Our audits were thorough, but there are still limits. As noted, any data not collected on a digital platform (e.g. RA annotations in Excel) require some trust. But even these platforms are vulnerable to implausible, though still possible, threats. For example, in studies such as Study 2 of CW-2019, it is possible that an unscrupulous researcher could register a large number of Craigslist accounts across the country, post fake ads selling iPhones, and rig responses to support a hypothesis. Likewise, for the studies that used crowd workers, someone could create many fake accounts and enroll in our surveys as a participant. The designs of our studies had several features that mitigated this risk. First, large sample sizes would have required fraud of a similar scale, which is hard to do. Second, most studies required participants to write open text, which requires more effort to fake. Third, in studies that only asked for Likert scale items (e.g. CR-2020 S1; CW-2019 S4A), participants were totally blind to treatment conditions. We believe the risk of such industrial-scale data fabrication is not high, nor is it specific to our papers - indeed, similar risks apply to most behavioral science research.

Discussion

The most important lesson we learned from our audits is that credibility is easier to plan for than to recover after the fact. The audit was much less taxing in places where documentation was built into the design of the research workflow - a fact that is also true of other quality control processes (Munafo et al., 2014). With that in mind, we offer recommendations at each step of a paper's journey, for how researchers can prioritize credibility moving forward. These are summarized in Figure 1.

[INSERT FIGURE 1 NEAR HERE]

Design Considerations

We suggest several ideas for the initial design of the research to improve credibility:

Collect data digitally. Digital data collection keeps a clearer record of the original measures. Although this may not always be possible, in most cases it should be prioritized. For example, paper and pencil surveys (which were used to assign pairs in Study 2 of CR-2020) are often easy to replace with online surveys instead. When direct digital data collection is not possible, digital records could be created in parallel - for example, if data are recorded by hand, timestamped photos of the recorded data could be taken regularly. Another common analog source is when data come from researchers themselves (e.g. annotations). In these cases, the cleaning code should be explicit about what the researchers see and do. The exporting and importing of data should be written into the code, indicating what was shown to the researchers, and how their data are merged back into the computational workflow.

Minimize use of PII. Every use of PII makes it harder to share the data after the fact. This information should be collected and used only when absolutely necessary. In Study 2A of QA-2017, we used participants' Mechanical Turk IDs to remove duplicates, using a dataset from

Mechanical Turk. We then used the IDs to link the ChatPlat transcripts to participants' Qualtrics surveys. These choices meant that extra raw datasets were required to completely reproduce the analyses, and other raw datasets could not be posted due to PII concerns. Another non-PII unique identifier could have easily been used to link datasets instead, and duplicate filtering is better done through the platform recruitment itself, rather than after the fact. These recommendations not only improve the transparency of the research, but also have the benefit of protecting participants from unforeseen harm due to data leakage.

Analysis Considerations

After data are collected, we have several recommendations for analysis:

Clean data with code. Researchers should not rely on human input to perform basic data cleaning tasks, such as removing extraneous rows and columns, sorting and pivoting data tables, or merging datasets. Computational pipelines are more reliable and reproducible, and errors are easier to debug. Doing these tasks by hand is sometimes seen as expedient. However, the time cost of learning coding techniques is paid back in productivity across many future projects. Efficiency is gained with practice, and old code can be repurposed for new data (by oneself and by others). Ideally, the cleaning code should be part of a preregistration for the study, so that cleaning choices are made while blind to their effects on the outcomes (Nosek et al., 2018).

Code together. Our workflow for QA-2017 and CW-2019 was not ideal - the task of entering results into the manuscript was conducted by different co-authors, who in some cases worked independently on their respective sections with only some double-checking of one another's work. Instead, we urge researchers to follow our approach in CR-2020, where every result was put into code and checked by another researcher. Even this did not guarantee

perfect reproducibility, but the decrease in differences in the number of discrepancies was apparent. We also recommend that after posting data, researchers then re-download the data and reproduce their results from scratch (ideally on a different computer). This is a stronger test of software dependencies (i.e. which version of a package produces the exact results), and will also uncover other issues that outside users may encounter. For solo-authored papers, researchers could exchange data and reproduce one another's work.

Writing Considerations

When the manuscript is being prepared, we also recommend the following:

Sweat the small stuff. What seems obvious and small to researchers deeply immersed in the analysis may not appear that way to outside readers, or even to those same researchers in the future, when their memory of the analysis pipeline has faded. No part of the analysis pipeline should be assumed to be clear from the written description - even trivial-seeming steps, such as rounding, setting random seeds, or specifying the package version, should be explicit in the code. Likewise, even secondary or tertiary results from the manuscript (e.g. reliability calculations, p-values from robustness checks) should be written into the analysis code.

Create a data sourcing worksheet*.* Authors and readers would be better off if all manuscripts were accompanied by a data sourcing worksheet. Such a worksheet would list the source for every step of data collection/processing (including cleaning, annotating, etc.) to simply trace the origins of each of the ingredients for the final paper. This step is easy to do when the manuscript is being written, but will only get harder over time. Taking this step will remove room for doubt and suspicion later on, and pre-empt the kind of ambiguity that allows bad actors to flourish.

Incomplete records are better than none. Several studies audited here had inadvertent gaps in their records, but we strongly believe that incomplete records are better than none. Rarely will an audit be truly impenetrable to fraud by a sufficiently determined researcher. Integrity gaps are unfortunate, but they are not evidence of fraud. Any efforts towards workflow documentation should be considered an improvement over the current status quo.

Audit Considerations

Finally, after a paper is submitted, we recommend the following for the auditing process:

Self-audit first*.* Even completely reproducible workflows may decay after posting, due to software updates, links to external data, and so on. The original authors will be able to solve these issues much faster on their own, before an external auditor gets involved. Furthermore, this process encourages authors to scrutinize their workflows with fresh eyes. They naturally suffer from a curse of knowledge when they submit the manuscript: what is obvious to them is not obvious to outsiders (Camerer, Loewenstein & Weber, 1989). Auditors and authors should also plan to be in contact (even anonymously, as we did for some of our reports). A collaborative process is much more efficient, because confusion can be resolved quickly.

Auditing is interdisciplinary*.* Some technical skill with data science is necessary for auditing. However, we found that domain knowledge was perhaps even more essential for efficient auditing, requiring that auditors be fluent in routine data collection and analysis techniques relevant to the research at hand. The best auditors will already be familiar with both, though any paper may still require auditors to learn some new skills on the job. The more familiar an auditor is with the relevant experimental protocols and analytical tools, the better.

Policy Considerations

For many researchers, these tasks may not be an enjoyable part of conducting research (Collins, Whillans & John, 2021), as they require large amounts of time, attention, and, at times, emotional resources. This happens in part because integrity is so rarely discussed, which leads to any mention of integrity - by readers or co-authors - being seen as an implicit accusation. We would like to lower the stakes around integrity documentation, and make it a more normal part of the scientific workflow. If these concerns are normalized, they can be addressed in advance, when it is least costly to do so.

But we are also realistic. Authors may only do this work if institutional actors make it clear that integrity is a priority. The incentives against bottom-up reform are strong. Researchers who voluntarily care about integrity take on an added burden, relative to those who do not. This will slow down their productivity, making them less desirable collaborators, reviewers, and colleagues. Furthermore, researchers who offer no transparency can assert perfection without evidence, while those who offer transparency expose themselves to critics.

It is clear there are not enough resources to conduct an independent audit for every published paper. Reproducibility is challenging enough to audit at scale, as Fisar et al. (2024) have shown. Integrity audits are even harder, because they often require communication and time to arrange privileged access to data. And some work, especially research conducted many years ago, is so hard to track that auditing for integrity would be nearly impossible.

We concede that many papers will not be audited, and yet we believe all papers should be *auditable*. The prospect of future research credibility cannot rest solely on retrospective audits. Instead, researchers should pro-actively demonstrate commitment to credibility during the study design phase, by documenting their data sourcing, cleaning, and analysis throughout their research. This should be further encouraged by policy changes from institutional actors. Journals editors, funders, and hiring committees should require data sourcing disclosure statements from researchers, including documentation of all data cleaning and attributions of which co-authors were responsible for each step. There could also be reputational incentives to

encourage integrity norms, such as badges or awards for researchers willing to audit others' work and/or be audited themselves.

Data collection platforms should also respond by making audits easier. For example, platforms could allow users to grant direct API access to raw data (minus PII), without requiring the researcher's account login information. They should also take greater effort to make researcher edits to raw data easily traceable. Google Sheets handles this well - it allows view-only access links, and tracks every edit to a document. iDecisionGames keeps internal logs of edits but these are not easily accessible - we had to request a bespoke investigation to check for changes. Qualtrics keeps records of edits to individual fields, but not wholesale deletions of observations. Better still, these platforms could limit the ability to edit data - ChatPlat only preserves raw data, requiring researcher edits to be done off the platform.

Conclusion

Science regularly requires evaluations of trust. But when objective systems for that evaluation are not in place, trust will instead be evaluated by subjective criteria, such as prestige, perceived guilt by association, or rumor (Mongeon & Lariviere, 2016). While trust is essential for the academic mission, it can lead to harm when it is given unwisely (Yip & Schweitzer, 2015). Furthermore, as the 2023 case has reminded us, no researcher can fully anticipate what scrutiny they will come under in the future, warranted or not. The best time to invest in credibility is now.

Acknowledgements

This work would not have been possible without the considerable efforts of our three independent auditors - Reny Baykova, James Zhou, and Nick Beazley-Long. All three were meticulous and diligent throughout the project. Additionally, we are grateful for financial support for the auditors' time from the Business School and Data Science Institute at Imperial College. We also wish to thank Marcus Munafò and the UK Reproducibility Network, which allowed us to find some of the auditors. Finally, the paper benefited greatly from comments suggested by Max Bazerman, Nicholas Brown, Frances Chen, Charles Dorison, Francesca Gino, Ariella Kristal, Don Moore, Marcus Munafò, Markus Perkmann, Uri Simonsohn, and others.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97(5),* 1232-1254.

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020, July). ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Collins, H. K., Whillans, A. V., & John, L. K. (2021). Joy and rigor in behavioral science. Organizational Behavior and Human Decision Processes, 164, 179-191.

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L. J., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of *Psychological Science*. *Psychological Science*, *34*(4), 512–522.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A. I., & Management Science Reproducibility Collaboration. (2024). Reproducibility in Management Science. *Management Science*, *70*(3), 1343-1356.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science, 3(4), 456-465.

Hardwicke, T. E., & Vazire, S. (2024). Transparency is now the default at Psychological Science. *Psychological Science, 35(7),* 708-711.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics*, *38*(2), 257-279.
Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, *113*(3), 430.

iDecisionGames. (2024). https://idecisiongames.com/promo-home

Jeong, M., Minson, J., Yeomans, M., & Gino, F. (2019). Communicating with warmth in distributive negotiations is surprisingly counterproductive. *Management Science*, *65*(12), 5813-5837.

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.

Mongeon, P., & Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers. *Journal of the Association for Information Science and Technology, 67(3),* 535-542.

Munafò, M., Noble, S., Browne, W. J., Brunner, D., Button, K., Ferreira, J., ... & Blumenstein, R. (2014). Scientific rigor and the art of motorcycle maintenance. *Nature biotechnology, 32(9),* 871-873.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour, 1(1),* 1-9.

National Academies of Sciences, Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, ... & Replicability in Science. (2019). *Reproducibility and replicability in science*. National Academies Press.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology*, *73*(1), 719-748.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115(11),* 2600-2606.

Nüst, D., & Eglen, S. J. (2021). CODECHECK: an Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. F1000Research, 10.

Ranganath, R., Jurafsky, D., & McFarland, D. (2009, August). It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 334-342).

Reis, H. T., Clark, M. S., & Holmes, J. G. (2004). Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness. In *Handbook of closeness and intimacy* (pp. 211-236). Psychology Press.

Simchi-Levi, D. (2023). From the Editor: Reflections on the Last Six Years as Editor-in-Chief of Management Science. *Management Science, 69(12),* 7151-7157.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7(6),* 670-688.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.

Yeomans, M., Brooks, A. W., Huang, K., Minson, J., & Gino, F. (2019). It helps to ask: The cumulative benefits of asking follow-up questions.

Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal*, *10*(2).

Yeomans, M., Minson, J., Collins, H., Chen, F., & Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, *160*, 131-148.

Yip, J. A., & Schweitzer, M. E. (2015). Trust promotes unethical behavior: Excessive trust, opportunistic exploitation, and strategic exploitation. *Current Opinion in Psychology*, *6*, 216-220.

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018, July). Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1350-1361).

Figure 1 - Forward-looking guidelines for behavioral scientists to improve integrity, throughout their research workflow.

RESEARCH DESIGN	***Collect data digitally, if possible.*** Digital data is much easier to check for errors, and is often independently stored by third parties (e.g. survey software).
	Minimize use of Personally-Identifying Information (PII). Not only can it potentially harm participants, it also unnecessarily complicates data sharing.
DATA COLLECTION	
DATA ANALYSIS	***Clean data with code***. Code can be inspected for errors. Human copy-pasting cannot.
	Code together. You will catch one another's errors, and become better coders.
PAPER WRITING	***Sweat the small stuff.*** What is obvious to you is not obvious to others (or to your future self).
	Create a data sourcing worksheet. Write down the full workflow before you forget, and before any doubt or suspicion can take root.
	Incomplete records are better than none. Perfection is impossible. Improvement is always possible. Beware the false pride of denying your own limitations.
PAPER SUBMISSION	
AUDITING	***Self-audit first***. If you cannot completely reconstruct your own workflow, starting from scratch, on a fresh computer, then an outside auditor will have no hope.
	Auditing is interdisciplinary. All modern social science is computational, to some degree. Still, the best auditors will understand both the tools and the protocols; both how and why something is done.

Table 1 - List of all studies in the three papers from the audit.

Paper	Study	Data Source	Posted to OSF			
			Clean data & analysis	Raw data	Cleaning code	Pre-registration
2020	1A	Qualtrics	Y	Y	Y	Y
2020	1B	Qualtrics	Y	Y	Y	Y
2020	2	Paper Surveys & iDecisionGames	Y	A	A	N/A
2020	3	Public	Y	Y	Y	N/A
2020	4A	Qualtrics	Y	Y	Y	Y
2020	4B	Qualtrics	Y	Y	Y	Y
2019	1	Qualtrics	Y	A	A	N/A
2019	2	Google Workspace	Y	**	A	N/A
2019	3	iDecisionGames	Y	A	A	N/A
219	3S	Qualtrics	Y	A	A	N/A
2019	4A	Qualtrics	Y	A	A	N/A
2019	4B	Qualtrics	Y	A	A	N/A
2017	1	Qualtrics & ChatPlat	Y	**	A	N/A
2017	2A	Qualtrics, ChatPlat & mTurk	Y	**	A	N/A
2017	2B	Qualtrics	Y	**	A	N/A
2017	3	Shared Privately	Y	**	A	N/A

Legend:

Y = Already posted to OSF during the initial submission process.
N/A = does not exist
A = Added to OSF during this audit.
** = Raw data relies heavily on PII (mturk IDs) for merging, matching, etc. Instead, the raw data will be shared privately w/ audit team via google drive for verification.

Table 2 - Audit checklists

Integrity
1. Is the raw data available from its original source?
2. Can we confirm that the raw data source has not been tampered?
3. Does the sourced raw data match the raw data that is posted on OSF?
4. Are there any data sources that are not digital?
5. Is the cleaning code available?
6. Does the cleaning code produce the cleaned data from the raw data?
Reproducibility
7. Is the clean data available?
8. Is the analysis code available?
9. Do the analysis code and clean data produce the numbers in the main text?
10. Do the analysis code and clean data produce the tables (where applicable)?
11. Do the analysis code and clean data produce the figures (where applicable)?

Appendix A - Raw Data Sourcing

Here, we describe the means by which we provided access to the auditors to the raw data, as best we could. This was successful, to varying degrees, depending on the source. In some cases, we used other researchers' data, so we could confirm that it was not tampered with (or at least, not by us). In other cases, we used common digital platforms to collect data, which kept independent logs. In one study, we created a novel digital workflow to conduct a field experiment, using Google Workspace, which keeps activity logs. Finally, we also had some human data entry in our workflows, for which it is harder to definitively prove there was no misconduct, although we still took some preventative measures.

Archival Data

Two studies relied on archival data borrowed from other researchers. Study 3 of CR-2020 used the Wikipedia editors dataset from convoKit, which is publicly available (Zhang et al., 2017; Chang et al,. 2020). On OSF we provide Google Colab notebooks to show exactly how that dataset was downloaded, and where it is stored and maintained.

Study 3 of QA-2020 used the SpeedDate corpus, which is not publicly available because the transcripts contain PII. The dataset was originally shared with us via email by one of the original authors, Dr. Dan Jurafsky. We asked that he share it with the auditors directly, but by the time of the audit (which was at least 14 years after the data were collected) he had deleted his original copy. Instead, we sought (and received) his permission to forward his original email to the auditors, which included the dataset. As an additional check, we connected our auditor with another researcher who used the same raw dataset for their research, to confirm the data were identical.

Digital Platforms

The majority of the studies across all three papers were conducted on Qualtrics, and involved asynchronous interactions. For example, Study 1A of CR-2020 recruited some people to write arguments on controversial issues, and Study 1B recruited other people to evaluate those same arguments, which were piped into Qualtrics via a Python web service. In these cases, the raw data are all stored in Qualtrics itself. The auditors all had their own Qualtrics accounts, so they could be added as a “collaborator” through the platform, which granted them access to the raw data as recorded.

Studies with synchronous conversation data were slightly more complex. Studies 1 and 2A in QA-2017 were run using a hybrid approach - Qualtrics assigned people to conditions and collected individual measures, while an embedded window from ChatPlat was used within the survey to enable conversations. Study 3 in CW-2019 and Study 2 of CR-2020 were both conducted within iDecisionGames, an online platform for conducting negotiations research that assigns conditions, collects measures, and allows for conversation. In these cases, we gave auditors our usernames and passwords, so they could download the raw data.

These three platforms are common software for research - Qualtrics especially - and offer varying degrees of protection against manipulation. At one end, ChatPlat makes it impossible for users to modify any data stored on their server. Conversely, Qualtrics and iDecisionGames allow researchers to change data on the platform itself. Qualtrics keeps records of most modifications, and when raw data is downloaded users can explicitly request the pre-modification data, which is what all the auditors did. Unfortunately, we discovered during the process (through contact with customer support) that Qualtrics does not keep any record when a participant response is deleted. Finally, iDecisionGames does not offer any user-facing option to check for modifications to the data. However, after correspondence with the company founder, they were able to check the server logs to see if any modifications were made to the data from the user account after initial entry from the participant. They provided a formal

statement attesting that they could not see a record of any modifications to the initial data (see Figure A1).

A Novel Digital Workflow

Study 2 of CW-2019 was a field experiment, which used a Gmail account to request discounts from people selling iPhones on Craigslist across 15 major US cities over the course of a month. The entire contents of the Gmail account could be downloaded as an MBOX file using Google Takeout, and the outcome of each interaction - the language of the seller's email, and their willingness to accept our requested discount - was extracted using a cleaning script. As the study was conducted, every sent email was logged in a Google Sheets spreadsheet that contained each seller's unique email address as a cross-reference to the assigned condition, and the discount requested. Google Sheets records the entire history of edits, so any tampering would also be revealed in this record.

Human Data Entry

While most of our workflows were digital, there were several places where datasets used human data entry (i.e., someone from the research team typing into Excel). These procedures inherently introduce risks to integrity, both unintentional or purposeful, since the human judgment criteria are not explicit, and can only be scrutinized up to a point. Arguably, this trade-off is justifiable when a task is too complex for an automated method. Still some efforts can be taken to minimize integrity risks.

For example, the first two papers relied on humans to proofread text written by participants for spelling and grammar errors. However, in all cases, the original uncorrected text was preserved, and we can confirm the results are substantively identical when the uncorrected text is used for analysis instead.

In Study 2 of CW-2019, we requested discounts from iPhone sellers online, and used their response emails as our outcome measure. We wanted to classify their responses as either “reject” (no response, no deal, or re-iterate original price), “accept” (respond affirmatively to the proposed discount), or “counter-offer” (respond with a compromise price). We had one of the original authors extract these response categories, which was double-checked by another of the authors. Because we kept the raw email data, the auditor could confirm that the extracted responses matched the text of the emails.

In Study 2 of CR-2020, some of the experimental methods were not digital. All participants’ issue positions were collected via paper-and-pencil survey. Additionally, in the first wave of data collection, participants were matched by hand - a research assistant sifted through the paper surveys, finding pairs of participants who disagreed on one of the three issues. Not only is this not reproducible, it also mistakenly resulted in some pairs who actually agreed on all three issues. In later waves, the paper survey responses were entered into a spreadsheet by a research assistant, and then fed into a sorting algorithm that generated pairs more efficiently.

Human judgment was used most in QA-2017. In one case, a team of research assistants provided turn-level annotations of different question types. After training, each RA was given an Excel spreadsheet where they could enter their answers. In that case, we kept the original Excel file in a timestamped Dropbox folder. For another part of the paper, research assistants provided conversation-level annotations of “responsiveness” (Reis, Clark & Holmes, 2004). Again, they were given raw Excel files, and wrote their evaluations into the spreadsheets. After they were finished, the analysis code shows how the spreadsheets were re-loaded back into R and merged into the main dataset. In both cases, RAs were blinded to the condition of each participant - our code record shows that the Excel files they were given did not contain the condition labels.

Human judgment was also used in many parts of the original data cleaning pipeline. For example, unnecessary columns in the data were removed by a researcher in Excel, as were

rows that were not filled by participants (e.g. pilot data), and some columns were renamed. During the audit process, we decided to write new cleaning code to reproduce all of this human input, so that the new cleaning code could take in the raw Qualtrics survey data directly.

One aspect that could not be documented in the code was for Study 1 of QA-2017, which was conducted in the lab. When there were technical difficulties, or a participant left during the study, a research assistant in the lab made a note of the exclusion in an Excel file. Although we later used R code to remove these 14 excluded participants based on the notes, we could not have replaced the human input entirely.

Figure A1 - Letter from iDecisionGames



iDecisionGames

idecisiongames.com
support@idecisiongames.com

September 20, 2024

To Whom it May Concern,

We are reporting in regards to a request made by Prof. Michael Yeomans and colleagues in regards to a set of exercises conducted on our platform under the user account registered by Prof. Julia Minson. They asked to check whether any of the data collected in those exercises was modified in any way by the user after it was initially collected, including: deleted observations, or changes to participant responses or internal variables. The nine exercises we were asked to check have the following IDs:

5fecd652-d003-420b-bbda-69b41b7af04c
25d75bae-faf3-46ad-bc63-1da7a50177a4
d1c98ddc-5614-4468-95ca-d8be0d26a507
76cc5ee0-819e-4d23-99a4-52fb36f4b469
2ed3f67b-b087-47e3-a1da-e83f6a4affb4
fa09e900-7f38-4539-bc54-e59ff0c586e0
4349e631-f65d-4835-b4b7-2fc02ef387cc
4d63b814-cec9-40f9-b81a-9dec61eabec4
12eac33e-0c99-4170-aac3-85deccfffcdc

We have thoroughly checked our data logs. Note that access to surveys and associated data is restricted to our senior devops team and all data is maintained in a database with a full audit record of any changes. Based on these checks, we have come to the conclusion that the data has not been modified in any way by the user account. The data currently stored on the server accurately represents all of the data as it was collected for these exercises. If you have any further questions regarding this process, please contact us directly at support@idecisiongames.com

Sincerely,

Niraj Kumar

Niraj Kumar
Founder and CEO, iDecisionGames
niraj@idecisiongames.com

Appendix B - Audit Reports

The reports are shared online at https://osf.io/rymv8/. The division of labor for the six reports was as follows.

Paper	Integrity Audit	Reproducibility Audit
QA-2017	Zhou	Baykova
CW-2019	Zhou	Zhou
CR-2020	Beazley-Long	Baykova