UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Zogmaister, C., Vezzoli, M., Facchin, A., Conte, F. P., Rizzi, E., Giaquinto, F., Cavicchiolo, E., Fusco, G., Pegoraro, S., & Simioni, M. (2024). Assessing the Transparency of Methods in Scientific Reporting. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.121243

Methodology and Research Practice

# Assessing the Transparency of Methods in Scientific Reporting

Cristina Zogmaister[1][a], Michela Vezzoli[2], Alessio Facchin[1,3], Federica Paola Conte[1], Ezia Rizzi[1], Francesco Giaquinto[4], Elisa Cavicchiolo[5], Gabriele Fusco[6], Sara Pegoraro[1], Maura Simioni[7]

[1] Department of Psychology, University of Milano-Bicocca, Milano, Italy, [2] Department of Psychology, Catholic University of the Sacred Heart, Milano, Italy, [3] Department of Medical and Surgical Sciences, Magna Graecia University, Catanzaro, Italy, [4] Department of Human and Social Sciences, University of Salento, Lecce, Italy, [5] Department of Systems Medicine, Tor Vergata University of Rome, Rome, Italy, [6] Sapienza University of Rome and CLNS@SAPIENZA, Istituto Italiano di Tecnologia, Italy, [7] Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

Keywords: Psychology, Open Science, Transparency, Reproducibility, Meta-Science

https://doi.org/10.1525/collabra.121243

Collabra: Psychology

Vol. 10, Issue 1, 2024

Enhancing transparency in scientific reports is crucial to foster trust, facilitate reproducibility, and ensure the integrity of research, ultimately advancing the progress of knowledge and innovation. To devise strategies for enhancing transparency in scientific reports, the initial step is to assess the current state: to objectively measure the current level of transparency, identifying its shortcomings and associated factors, and to gauge improvements, for instance, following interventions. Here we present a new tool and a proof of concept to this endeavor.

Using a checklist, we evaluated the methods transparency of a corpus of 180 papers published in 2011 and 2021 in five top-tier psychology journals. We specifically focused on the materials, procedures, and characteristics of population and sample. We summarized the level of transparency in the methods of each paper with the Transparency Of Methods (TOM) score. This score consists of the proportion of relevant information regarding the method that is available to the reader of a scientific report, either in the main text of the paper, or the appendixes, supplementary materials, and online open repository. It ranges from 0 (i.e., no transparency in the relevant aspects of the methods and materials) to 1 (i.e., the scientific report is fully transparent in all the relevant materials and methods).

The results affirm TOM's potential for assessing the transparency of scientific reporting and offer two snapshots of transparency in the methods of published papers, a decade apart. While they indicate progress has been made, there remains room for further enhancements and highlight specific areas that require attention.

In conclusion, this work underscores the ongoing need for improvement in methods' transparency and introduces a valuable tool, demonstrating its applicability as a means to evaluate the transparency of scientific reports.

Transparency in reporting methods consists of clear and comprehensive documentation of the processes, techniques, and procedures employed during the study. It involves providing detailed explanations of the population and sampling procedures, the manipulations and measurements used, and other relevant aspects of the research process. A simple approach to assess the transparency of the reporting of the methodological aspects of a study could be, for example, to ascertain whether another researcher, upon reading the article and utilizing the materials provided by the author (e.g., supplementary materials online, links to repositories containing stimuli, etc.), would have all the necessary information to replicate the study from scratch, in a manner very similar to how the original study was conducted. If a scientist, based on the paper's reading and available materials, doesn't have sufficient information to conduct a replication, this means that the reporting doesn't describe the methods fully transparently. One could argue that when a paper lacks methodological details, readers can reach out to the authors and request additional information. However, this approach is far from optimal. Firstly, it is not always successful: As it will emerge from the present research, it is not consistently clear to whom one should address these inquiries; moreover, at times the corresponding author is no longer active in academia; finally, we are not aware of studies that specifically addressed sharing research materials upon request, but personal experience show that requests for shar-

a Correspondence should be addressed to Cristina Zogmaister; Email: cristina.zogmaister@unimib.it

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

ing materials may go unanswered or receive a negative response (e.g., due to technical issues like a broken computer or an office relocation) and studies on data sharing clearly indicated that the request for research data from authors in a substantial proportion of cases proves unsuccessful (see e.g., Stodden et al., 2018). Secondly, even when authors do answer the request and share the materials, it usually entails a time-consuming process.

This work was stimulated by our informal observation, which we shared with colleagues: When attempting to replicate a scientific experiment, the scientific report of the original study often lacks crucial information. Moreover, transparency in the methods is vital not only for replication. It is also a key preliminary requirement for understanding and evaluating research, expanding its results, and comparing studies. A transparent methods section might contribute to bolstering the overall credibility of the research.

## The Importance of Transparency in Comprehending and Appraising Empirical Research

For those readers who wish to gain a general understanding of a study, a definition of key concepts and descriptions of the core aspects of manipulations and measures may be sufficient. Even in such a case, however, reading the specific items that make up the questionnaire, for example, would allow a deeper understanding of the construct that is being measured. Seeing all the items used, instead of reading their summary description or viewing just one or two stimuli that the authors deemed particularly representative, can be highly informative. Since an example often clarifies more than a thousand words, let us imagine an experimental study that explores reaction times to sexualized and non-sexualized human bodies: The report of such an experiment might state that pictures of sexualized and non-sexualized human bodies were presented to participants to evaluate, and perhaps contain one image of a sexualized and one of a non-sexualized body. Making all those pictures available would allow the reader to better understand what is meant by 'sexualized' and 'non-sexualized' in this research. It would also help to identify other possible differences between the two sets of stimuli besides the level of sexualization and hence help catch confounding variables or methodological artifacts. The sexualized stimuli might inadvertently differ from the non-sexualized ones, for instance, in visual complexity (e.g., related to a higher amount of visual information in the clothing in the photos of less sexualized bodies, or more asymmetrical postures in the photos of the more sexualized bodies) and this complexity could impact on the reaction times. Systematic confounding differences between the two corpora of stimuli might not emerge by inspecting only a few sample items, but they could become clearer when examining the entire set of stimuli. Another scientific report might state that participants completed a scale of ingroup identification and present two of the scale's items as examples. Again, the whole set of items would provide a more complete idea of the operational meaning given to ingroup identification in the study.

A scientific report should detail all manipulations, experimental conditions, and measurements of the study. However, these aspects are not always exhaustively reported (Franco et al., 2016; John et al., 2012; LeBel et al., 2013) in the published literature. A statement in the manuscript explicitly stating that all the manipulations and measures are reported could have the double benefit of increasing readers' confidence in the absence of questionable research practices, such as cherry-picking (Suter, 2020), and serving as a "nudge," in that it might encourage some researchers to disclose all manipulations and measurements. Such a statement may have a limited impact in cases of fraud; however, we believe that actual instances of fraud constitute a tiny minority among researchers' behaviors (Agnoli et al., 2017; Stricker & Günther, 2019).

In sum, an increased transparency of methods would foster a deeper understanding for readers, editors, and reviewers. This, in turn, could contribute to more rigorous and reliable scientific discourse. Overall, allowing readers to better judge the quality of the research could increase the overall trust in the research (see Vazire, 2017).

## The Relevance of Methods Transparency for Replication Studies

The reasons why the replication of scientific studies is important have been extensively discussed elsewhere (e.g., Simons, 2014). Sharing the research materials and all relevant information facilitates accurate replications (e.g., Funder et al., 2014). An issue in replication studies lies precisely in the challenge of discerning whether discrepancies between the outcomes of a new study and those of the original stem from the original effect's lack of robustness (for example, the original effect might be unreliable, or due to a false positive), or from deviations in procedure from the initial research (Laraway et al., 2019; Leek & Jager, 2017) and contextual differences ("hidden moderators"; Van Bavel et al., 2016). For instance, a certain effect might be affected by participants' characteristics, that differ in the original and in the replication study, as for example socioeconomic status (Markus & Stephens, 2017). Lack of transparency in the description of the population and sampling methods could hinder the possibility of discovering this difference between the two studies. In such a scenario, the lack of methodological transparency would inadvertently impede scientific advancement: By possessing comprehensive information, the divergence between the two studies could spark the formulation of novel research hypotheses concerning the variations in effects based on population characteristics. However, due to the absence of such details, conclusions are necessarily limited to the acknowledgement that replication did not occur. Having access to a clear and transparent methods section, along with all stimuli, serves not only to enable more accurate replications but also to eradicate these sources of ambiguity when determining the success or failure of replication and the underlying causes (Grahe, 2018).

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

## The Role of Methods Transparency in Facilitating Comparative Studies

Besides failed replication attempts, there are many other instances in which apparently similar studies find diverging results. Also in such cases, if materials and methods are available, researchers can thoroughly analyze and scrutinize the fine details of stimuli and procedures employed in the different studies, of populations and recruitment methods, *et cetera*. Transparency of methods would enable the identification of potential factors contributing to the diverging results and the formulation of informed hypotheses about the underlying reasons behind these discrepancies, which could be subsequently tested.

## Methods Transparency and a More Efficient Use of Public Resources

Science is a collective endeavor. By giving colleagues access to the tools, materials, and methodologies utilized in an experiment, authors allow other scientists to build upon the existing knowledge to ask new questions and design new experiments. The possibility to reuse materials and methods, made possible by transparent reporting of methods, will further improve the efficiency of research: Many resources are invested in preparing the methods, stimuli, and scripts required for conducting scientific research. Having access to the materials that other scientists have developed can save time and other resources that would otherwise be spent on "reinventing the wheel," hence leading to more efficient use of public resources and accelerating scientific progress.

## Individual Benefits for Researchers who Report their Methods Transparently

Although we are not aware of any systematic research on this aspect, we believe that openly sharing one's methods and materials, and in general transparency of methods, could also lead to individual benefits for researchers. The enhanced comprehension of research facilitated by the transparent reporting of methods will bolster both the researchers' and their study's credibility, especially in the eyes of those editors, reviewers, and readers who are most sensitive to issues of transparency and robustness in scientific research. Moreover, those who utilize the shared materials and methods should attribute proper citations to the source article, thereby augmenting the visibility of the work of researchers writing transparent methods. Furthermore, some journals and funding bodies are now mandating the availability of open materials. Consequently, embedding practices of transparent methodology in one's scientific research flow aligns with these requirements and streamlines responses to these requests.

It is important to acknowledge that transparency also has downsides. Firstly, it should not be pursued at the expense of clarity. Overly comprehensive and complex disclosures could overwhelm the reader and be undigestible. To overcome this possible limitation, authors can distinguish the elements of primary importance - that should be reported in the article proper - from those that are potentially secondary and can thus be made available in supplementary materials without detracting from the clarity of the main scientific report. Moreover, the demand to make everything transparent might create stress, perhaps even more so for early-career researchers who may feel that every aspect of their work is under scrutiny. Despite the challenges and risks associated with method transparency, the benefits of transparent practices support the foundational principles of good scientific research and contribute to its objective of broadening human knowledge and solving complex problems. Finally, one might argue that in certain cases transparency might decrease the credibility of a research. This could happen for a poorly conceived and conducted study, where a wholly transparent report would evidence the mistakes and low value of the study. In a case like this, the benefit to society of being able to understand the quality level of the study would be in opposition to the immediate interest of the researcher, who would have the low quality of their work exposed. Therefore, in evaluating the pros and cons of scientific transparency, the personal values of the researcher come into play. For this reason, we believe reporting research methods transparently should not be left to the discretion of individual researchers but should become a shared practice and a common expectation. If that were the case, the additional efforts and commitment to transparency would be distributed fairly across all, rather than solely falling on those who are most committed to a more open, shared, and rigorous science. Moreover, it would be the absence of transparency, and not its presence, that would be a signal of atypical behavior.

## The Need of a Checklist for Assessing Methods Transparency in Scientific Reporting in Psychology

A common experience we share with other researchers who have tried to replicate published studies, adopt the procedure of another researcher for their own studies, or even just understand the details of an experiment, is that many published scientific reports fall short in delivering the comprehensive information necessary to fully comprehend an empirical study. While some papers feature very clear and informative method sections, link to an open repository containing all the stimuli, include verbatim instructions and a detailed description of the procedure, *et cetera*, others offer insufficient details. Occasionally, a paper may reference additional online materials, yet the functionality of the provided links is not reliable. Other times, the supplementary resources are accessible, but do not consistently encompass the entirety of the required information.

Having an accurate and comprehensive understanding of the situation is crucial. We need to go beyond the mere impression that scientific reports should be more transparent and gather a more objective overview of the landscape. Just to list a few key concerns on this issue: How much of their methods information do authors typically share in published reports, and perhaps more importantly, which information do they typically share (and which do they not)? Can we spot a trend of change towards an increased level

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

of transparency? Are there differences between sub-disciplines? Are there differences in the methodological transparency of the scientifically younger and senior authors? Do interventions to foster methods transparency (e.g., journal guidelines or funders' requirements) have an impact? Answering these questions would foster a deeper comprehension and be beneficial for driving impactful change.

In general, however, the prevalence of transparent reporting of psychological methods is largely unknown. There is empirical evidence regarding the practice of making research materials available: Hardwicke and colleagues (2022) examined 250 psychology papers, randomly sampled from articles published between 2014 and 2017, and found that 14% (26 out of the 183 empirical papers in their sample) contained a statement regarding the availability of original research materials. In the even lower number of 19 cases, the materials were indeed available, whereas in the other 7, they were not available due to broken links (see also Hardwicke et al., 2020, for similar results in a sample of articles published in the social sciences). Klein and colleagues (2012) analyzed the method sections of 346 experiments reported in two psychology journals, *Psychological Science* and the *Journal of Personality and Social Psychology*, focusing on some specific aspects of the procedure. They noticed that the presence/absence of the experimenter during the session was mentioned in approximately 30% of the studies, and the information on how the study was presented to participants was present in less than 40% of the studies. Besides this, we have found no systematic study regarding what information researchers typically make available regarding their methods, and what information, to the contrary, they often fail to report. Therefore, with the present work we want to offer a contribution towards this goal, and we specifically focus on the materials, procedures, characteristics of population and sample.

We reasoned that the ideal instrument to assess transparency of methods should provide two types of information. Firstly, it should reliably assess the availability of specific methodological information in scientific reports, so that we would be able to understand what types of information researchers usually report and what other types they usually do not report. In turn, this could allow us to reflect on the reasons why researchers typically neglect specific information and investigate whether there are differences across subdisciplines, just to give a few examples. Secondly, such an ideal instrument should provide an overall score of methods transparency of the scientific report, to address questions such as whether we can see an increase in transparency in scientific reporting today, as compared to the situation a decade ago, before open science became a central topic for the psychological community. This work is a feasibility study (or a "proof of concept") for such an instrument.

Regarding this instrument, the first requirement is a structured checklist of the essential components that a scientific report must encompass. Luckily, when we initiated this project, the PECANS initiative (PECANS, 2024) was underway, and the first round of dimensions was already available.

### *The PECANS checklist*

The PECANS initiative, where PECANS stands for "Preferred Evaluation of Cognitive And Neuropsychological Studies," aims to provide a consensual checklist of the methodological pieces of information that a scientific report in cognitive and neuropsychological research should include. The PECANS checklist employed the Delphi method, a widely used technique aimed at establishing a reliable consensus among a panel of experts across various scientific fields (Barrios et al., 2021; Keeney et al., 2006). This survey approach includes multiple rounds, reiterated until stability in responses is achieved. Anonymity among experts is ensured by using written questionnaires, and the process involves controlled feedback provided by a facilitator after each round. Additionally, statistical aggregation of group responses after each round is provided to offer feedback to participants (Barrios et al., 2021; Rowe & Wright, 1999; Trevelyan & Robinson, 2015; von der Gracht, 2012).

When we started this project, the Delphi study was not yet completed. An international group of scientists had created an initial list of items that they considered important for the transparency, quality, and reproducibility of an experimental work and 206 international experts had indicated which items, among those listed, should be included in a scientific report; the experts were also asked to provide additional suggestions on how to improve the checklist. Although the PECANS checklist has a specific focus on studies in cognitive psychology and neuropsychology, an inspection by our research team indicated that many of its items could be easily adapted to experimental psychological research in general. Therefore, we developed the measurement instrument used in this study, which is described in the Methods section (see also Table 1), using the PECANS work as our starting point. We discarded the PECANS items that we deemed not directly related to the objectives of the current goals (e.g., those concerning the theoretical rationale, whether the hypotheses were preregistered, and the statistical aspects). Indeed, empirical research is an intricate and multifaceted process, warranting a phased approach to thoroughly examine each aspect without undermining others. We excluded preregistration from our study because we reasoned that reducing the issue of its presence/absence would be inadequate, and a more appropriate investigation would involve the analysis of what had been preregistered (e.g., hypotheses, materials, procedures, etc.), any deviations from preregistration to study, and whether these had been reported, hence demanding a standalone effort. Moreover, while preregistration of materials enhances transparency, it is not essential for understanding in detail how the study was conducted. Similar considerations apply to data analysis aspects, including power analysis and effect size justification. These are crucial and rich in informational value, meriting independent exploration. Machine learning procedures may address these aspects more efficiently, and they are also subject to scrutiny by other researchers. We next removed redundant items in the

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

PECANS checklist (i.e., items that described the same piece of information with different phrasing), and adapted and restructured the items to the goals of the present instrument to assess the methods transparency (e.g., at times, one PECANS item was divided in two or more items in the Transparency Of Methods (TOM) checklist).

Finally, we added six new items, that were not present in the PECANS checklist: Concerning the characteristics of the population and the sample, element that was present in PECANS (item 6. Population of reference and related exclusion conditions), we added items 7 and 8, which further specified the sampling procedures, and item 24 regarding the information provided at the invitation to participate in the study, to better qualify the sample drawn from a given population. We deemed that information provided at the invitation be important also for their potential influence on expectations and potential demand characteristics. We added item 23 regarding manipulation checks, because of the potential impacts on results (Ejelöv & Luke, 2020), and item 38 concerning filler tasks for the same reason (e.g., Arehalli & Wittenberg, 2021). Finally, we added item 29 (availability of the experimental script) because, similar to the script of the statistical analysis, it can provide an unambiguous description of the fine details of a study.

## The Present Research

Here we assessed the feasibility of using the checklist in Table 1 to investigate the transparency of methods reporting in psychological research. The checklist concerns the presence vs. absence of specific information in a scientific report. The raters' task is to judge, for each item, whether the scientific report provides the information described (either in the main text, appendixes, supplemental materials, or an open repository linked in the scientific report). To help resolve possible doubts, the raters are instructed to ask themselves whether the information in the scientific report would be sufficient to replicate that aspect of the study exactly, or not. For each item, they have to assign a score of 1 if the information is present, 0 if it is absent, or indicate that the item is irrelevant to the scientific report at hand. Based on the evaluation of a report on each item, an overall Transparency Of Methods (TOM) score can be computed as the mean of the 1 and 0 scores. Therefore, the TOM score can vary between 0 (none of the relevant items are reported) to 1 (all the relevant items are reported) and indicates the proportion of relevant methods information contained in the scientific report. Three key elements are central when evaluating a measurement instrument: usability, validity, and reliability.

Concerning the *usability* of the list of dimensions, we addressed how much time would be required, on average, for a rater to evaluate each scientific report. We reasoned that, to be useful in the measurement and analysis of transparent method reporting, the checklist would need to be relatively quick to compile, even for a rater evaluating an unfamiliar piece of research. Perhaps in the future an automated Artificial Intelligence system could be used to perform such a task, but as long as human raters are required, time-efficiency will be of the essence. Thus, one of the aims of this research is to gather information on whether the checklist could be used in an easy and efficient way to rate the level of transparency of the methods in a scientific report.

Validity is critical to selecting and applying measurement instruments, ensuring that they accurately capture the targeted constructs. While various types of validity are essential, two types directly impact the credibility and usefulness of the measure presented in this study: content validity and construct validity. Content validity is essential as it ensures that the selected items are relevant and accurately represent the target construct (Almanasreh et al., 2019). This evaluation typically involves using expert panels to assess items based on their relevance and representativeness to the content domain (Almanasreh et al., 2019). To comply with this definition, the present measure of method transparency derives its content validity from being grounded in various experts' judgments: the judgments of the group of scientists who developed the initial checklist with the goal of listing the elements that a scientific report should contain to increase the transparency and reproducibility of scientific reporting, those of the second group of experts who, participating to the Delphi study, assessed the relevance of the set of items, and finally the judgements the present group of authors, who assessed the representativeness of the content domain. On the other hand, construct validity is indispensable for establishing the theoretical foundations of measures (Cronbach & Meehl, 1955). The traditional conceptualization of construct validity emphasizes establishing significant relationships between test scores and external variables, such as other measures or criteria known to be related to the construct (Campbell & Fiske, 1959; Kimberlin & Winterstein, 2008). Given the crucial importance of this dimension for the use of a measure, we decided that, in this initial feasibility study, we could gather preliminary evidence of construct validity based on two expectations. Firstly, we reasoned that there should have been an increase in the overall level of transparency in the reporting of the methods from papers published in 2011 to 2021 because of the heightened awareness within the scientific community regarding the importance of replicability and openness in research. Thus, we compared the TOM scores of two corpora of scientific reports, one from 2011 and another from 2021, and hypothesized that the average score would be higher in the more recent papers. We also expected that scientific outlets with a heightened awareness of reproducibility and open science concerns would publish more transparent papers. We specifically choose a journal highly connotated in terms of attention to the openness of reporting. Also in this case, a higher average TOM score for the papers published in this

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

journal would be a positive indicator of the construct validity of the measurement instrument[1].

Concerning the reliability of the measure, the information listed in Table 1 pertains to specific details: Therefore, one might expect their presence in a scientific report to be objectively evaluated. However, a piece of information can be reported with different degrees of detail, or explicitness, and this brings a degree of subjectivity in judgment. We found it crucial to examine the level of subjectivity in determining their presence or absence. Additionally, we aimed to determine if this subjectivity could be minimized through clear instructions and adequate training for raters. Therefore, we ascertained the interrater reliability of the measurement for the various items and the overall TOM score. Based on the experience that the first group of five raters had with the checklist in Study 1, we improved the instructions and in Study 2 two new raters reassessed a sample of the papers, to test the interrater reliability of the instrument with the novel instructions. Finally, in Study 3 we investigated whether enhanced training could improve the reliability of the TOM scores.

## Focus and Setting of the Study

The checklist could be potentially used for various types of empirical studies. For this proof of concept, we narrowed down the investigation to experimental research, because this would allow a meaningful comparison between two corpi of materials, eliminating the possible variability due to differences between types of research (e.g., studies with and without experimental manipulation of variables) and therefore keeping the overall number of manuscripts within reasonable levels given the available resources (i.e., number of raters). For the same reason of allowing meaningful comparisons while keeping the corpus of materials manageable, we chose five target journals. As the Open Science Collaboration (2015) evidenced that the replication rate was lower in studies published in journals representing social psychology (25%) than for cognitive psychology (50%), we deemed it particularly interesting to investigate the level of methodological transparency in social psychology scientific papers. It is plausible that this very field has undergone a distinct reflection into the issue of open science and reproducibility practices, and therefore an inspection in this specific domain would provide a good test for our hypothesis of an increase in transparency of methods. We, therefore, chose two target journals publishing research in social psychology: the *Journal of Personality and Social Psychology* (JPSP) and the *European Journal of Social Psychology* (EJSP). As the JPSP publishes papers in three sections: *Attitudes and Social Cognition*, *Interpersonal Relations and Group Processes*, and *Personality Processes and Individual Differences*, we focused our research only on the first two sections of the journal. We chose two other top-tier journals publishing experimental research in areas different from social psychology to test our checklist also on a more diverse set of articles: *Cognition*, a journal in cognitive science, and the *Journal of Experimental Psychology: General* (JEPG), a more generalist journal publishing research of interest for different psychology communities[2]. Finally, we analyzed papers published on *Collabra: Psychology* (hereafter, Collabra); this being the official journal of the Society for the Improvement of Psychological Science, we expected a high level of sensitivity to open science and awareness of the methods transparency needs from the Editorial Board, reviewers, and authors of this journal. We deemed it, therefore, interesting to check if papers published in this journal would be characterized by greater methods transparency as compared to the other selected journals. As Collabra did not exist in 2011, we analyzed only the papers published in 2021 for this journal. Collabra consists of various disciplinary sections; for reasons of comparison with the other journals, we selected specifically the two sections of social and cognitive psychology.

The rationale behind this journal selection decision, which implied that our articles would not constitute a representative corpus of psychology articles published during the two years of interest, was that by selecting four specific journals, all with a distinct focus on experimental research, the type of articles and content covered would not dramatically differ from 2011 to 2021. This made comparisons more feasible and meaningful. Metaphorically, the intention was akin to capturing two snapshots of a landscape a decade apart to discern what changes had occurred over this time.

In sum, our study is different from previous works that also investigated the transparency of methods (Hardwicke et al., 2020, 2022), in that we provide a more focused picture, that encompasses a specific type of scientific reports, those of experimental psychological research in social and general psychology, while Hardwicke and colleagues had a broader focus on psychological research (2020) and the social sciences (2022). This will allow us to provide more fine-grained information about this specific area of scientific research. Our work is also different from Klein and colleagues' (2012) because we investigated a substantial group of items of information that span the various aspects of the reporting of the methods section of an experimental study in psychology. Also, we took two separate snapshots of methods transparency, one before (2011) and the other after (2021) the issue of reproducibility in psychological research came to the forefront of scientific debate. Our endeavor serves a dual purpose: On the one hand, it tests the viability of the method we had conceived – that is, whether an evaluator can render an objective and reliable assessment of the transparency level within an article's method-

1 Note, however, that we did not pre-register this second hypothesis because we could locate only one journal with a specific connotation in terms of open science.

2 The JEPG "covers research that is of broad interest or bridges the traditional interests of two or more communities of psychology" (from the Journal scope statement, https://www.apa.org/pubs/journals/xge).



Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

ology with a reasonable consumption of time resources – on the other hand, it tests the construct validity of the derived values. If the tool proves effective, it will provide us with an instrument to gather objective insights into the methodological transparency level of psychological articles.

## The Checklist

Our checklist was built upon the efforts of the PECANS initiative (PECANS, 2024). We considered the items that emerged from the initial phase of the PECANS initiative and, drawing extensive inspiration from this work, identified the five categories and the 48 Transparency Of Methods checklist (TOM) elements, listed in Table 1.

The first category, general requirements, consists of five items concerning general information about the scientific report, specifically whether it explicitly states the inclusion of all experimental conditions and dependent variables (a key element to assure the reader that questionable practices like cherry-picking have not been employed), the presence or absence of open materials and, in cases where contacting the authors for materials is necessary, whether the procedures for obtaining these materials are adequately clarified. Regarding the general requirements, Item 1 warrants an explanation. One could argue that the crucial aspect for methodological transparency is not the presence of a statement affirming that all measured variables and conditions are reported, but rather the fact that these pieces of information are actually provided in the report. However, without this explicit statement, the reader cannot know whether this is the case. Questionnaire data (e.g., Fiedler & Schwarz, 2016; John et al., 2012) and objective measures (Franco et al., 2016), indeed, provide empirical evidence of selective reporting of measures and experimental conditions.

The second category, consisting of 11 items, pertains to information about the experiment's participants: the reference population and exclusion criteria, the recruitment and compensation strategy, whether it is clearly specified how many individuals were invited to participate and how many were included in the analyses, demographic characteristics, and how they were assigned to experimental conditions.

The subsequent research design category is composed of six items that concern the type of blinding (if any), the distinction between manipulated and measured variables, the methods by which they were operationalized and whether this was done within or between participants, the counterbalancing of materials, control conditions, and manipulation checks.

Next comes the procedure category, which comprises 14 items. These pertain to the type of information participants received when invited to participate, whether the study was conducted online or in-person, the number of sessions comprising the study and its duration, task sequencing, filler tasks, and characteristics of the setting, including both software and social setting. Not all items are relevant to every study; for instance, some only apply to online studies, while others are specific to in-person studies.

The final category, consisting of 12 items, specifically pertains to materials: tasks, instructions, stimuli, and response modalities.

# Method

## Design and Open Practices Statements

This was a retrospective observational study with a cross-sectional design. Sampling units were individual articles. The measured variables are described in Table 1, first column. We preregistered the study (https://osf.io/4vxgc). We report all deviations from the preregistration in the "Deviations from Preregistration" Section. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data, materials, and analysis scripts related to this study are openly available on OSF (https://osf.io/etkh5/).

### *The Sample*

We obtained a sample of 180 psychology articles describing experimental studies published in the year 2011 (80 papers) and 2021 (100 papers).

For JPSP, EJSP, Cognition, and JEPG, we selected the first 20 papers, in chronological order from the first issue of 2011, that described at least one experimental study; we also selected the first 20 papers describing at least one experimental study published in 2021. For Collabra, we selected the first 20 papers describing at least one experimental study published in 2021. For the two journals (JPSP, Collabra) that had two sections of interest, we selected, within each year, an equal number of papers for each section. We used the following inclusion criteria: The manuscript had to be a regular paper (i.e., no short reports, or "fast track" reports in the case of EJSP); it described at least one experimental study (i.e., a study in which at least one independent variable was manipulated); the unit of analysis consisted of humans (i.e., studies conducted on animals and simulation studies were discarded). If the paper described several studies, we coded the first experimental study. The list of DOIs of the papers is provided on OSF.

The number of papers was chosen based on the available resources.

## Procedure

The papers were classified on the above dimensions between August 7 and September 24, 2022. The materials were coded by five raters (AF, ER, FC, FG, and MV). Each rater was randomly assigned 44 papers (i.e., 24.4% of the papers) so 40 papers were coded by two coders to examine inter-rater reliability. For each manuscript, the rater evaluated the first experiment described and categorized each feature reported in Table 1 as present (1), absent (0) or not relevant (9). Each feature was considered present if the rater could find it in the main paper, appendixes, supplemental materials on the journal site, or other linked materials in an open repository. The coders carefully read the methods sections of the first experimental study in

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

**Table 1. Features evaluated in the TOM checklist and percentage of papers evaluated as reporting each information, considering all papers and depending on the year of publication.**

| Checklist items | Reliability | Publication Year 2011 (n = 80) | 2021 (n = 80) | Collabra (n = 20) |
|---|---|---|---|---|
| **General (minimal) requirements** | **ICC = .51** | | | |
| 1. Statement that all conditions are reported. | A = .65<br>κ = .00 | 15<br>(18.75%) | 24<br>(30.00%) | 3<br>(15.00%) |
| 2. Statement that all D.V.s are reported. | A = .62<br>κ= -.11 | 15<br>(18.75%) | 24<br>(30.00%) | 3<br>(15.00%) |
| 3. Presence of open materials containing additional information regarding the Methods. | A = .72<br>κ = .45 | 9<br>(11.25%) | 45<br>(56.25%) | 17<br>(85.00%) |
| 4. In the case of materials available upon request, specify exactly what conditions need to be met in order to obtain the materials or make it clear that the materials will be released unconditionally. | A = .93<br>κ = .76 | 1<br>(1.75%) | 15<br>(31.91%) | 4<br>(50.00%) |
| 5. A clear indication of to whom requests for materials are to be directed. | A = .79<br>κ = .51 | 5<br>(8.77%) | 11<br>(23.40%) | 6<br>(75.00%) |
| **Participants** | **ICC = .68** | | | |
| 6. Population of reference and related exclusion conditions, e.g., students, workers, US citizens. | A = .75<br>κ = .28 | 51<br>(63.75%) | 69<br>(86.25%) | 15<br>(75.00%) |
| 7. Sampling method (clearly reported or understandable from description; e.g., convenience sampling). | A = .50<br>κ = .00 | 25<br>(31.25%) | 38<br>(47.50%) | 10<br>(50.00%) |
| 8. Recruitment strategy, e.g., university pool, leaflets -where posted?-, posts on social media. | A = .65<br>κ = .24 | 38<br>(47.50%) | 55<br>(68.75%) | 15<br>(75.00%) |
| 9. Compensation, e.g., course credit, money, none. | A = .75<br>κ = .47 | 49<br>(61.25%) | 49<br>(61.25%) | 13<br>(65.00%) |
| 10. Method for assignment to condition, e.g., random. | A = .66<br>κ = .27 | 49<br>(65.33%) | 46<br>(63.01%) | 15<br>(75.00%) |
| 11. Demographics (gender, age). | A = .78<br>κ = .42 | 37<br>(46.25%) | 66<br>(82.50%) | 16<br>(80.00%) |
| 12. Demographics ethnicity descriptives. | A = .85<br>κ = .70 | 14<br>(17.50%) | 26<br>(32.50%) | 5<br>(25.00%) |
| 13. Demographics education level. | A = .92<br>κ = .54 | 11<br>(13.92%) | 11<br>(13.92%) | 2<br>(10.53%) |
| 14. Other demographics, e.g., medications, diagnostic criteria (where relevant). | A = .56<br>κ = .01 | 19<br>(25.68%) | 21<br>(29.58%) | 5<br>(27.78%) |
| 15. Indication of how many participants were contacted and how many were included in the analysis. | A = .66<br>κ = .29 | 13<br>(16.25%) | 17<br>(21.25%) | 4(20.00%) |
| 16. Indication of whether the study has a blind design, e.g., single-blind, double-blind. | A = .62<br>κ = .24 | 30<br>(38.96%) | 26<br>(35.13%) | 9<br>(47.37%) |
| **Design** | **ICC = .30** | | | |
| 17. Description of which are the manipulated variables, and the measured variables. | A = .85<br>κ = .41 | 66<br>(82.50%) | 71<br>(88.75%) | 18<br>(90.00%) |
| 18. Description of operationalizations in sufficient detail. | A = .70<br>κ = .15 | 62<br>(77.50%) | 64<br>(80.00%) | 17<br>(85.00%) |
| 19. Specification of whether the independent variable is manipulated between or within participants (for each manipulated variable). | A = .78<br>κ = .05 | 64<br>(81.01%) | 70<br>(89.74%) | 19<br>(95.00%) |
| 20. Indication of whether materials are counterbalanced between participants. | A = .69<br>κ = .38 | 25<br>(32.05%) | 30<br>(40.00%) | 14<br>(73.68%) |
| 21. Description of control conditions (what does the control group do? e.g., active/passive controls). | A = .66<br>κ = .32 | 36<br>(46.75%) | 40<br>(56.34%) | 12<br>(66.67%) |

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

| | | | | |
|---|---|---|---|---|
| 22. Description of manipulation checks. | A = .84<br>κ = .52 | 21<br>(26.58%) | 31<br>(38.75%) | 9<br>(45.00%) |
| **Procedure** | **ICC = .68** | | | |
| 23. Description of the information participants are provided at the invitation to participate. | A = .62<br>κ = .12 | 25<br>(31.25%) | 24<br>(30.00%) | 9<br>(45.00%) |
| 24. Indication of where the study took place, e.g., online, in the lab. | A = .68<br>κ = .33 | 45<br>(56.25%) | 54<br>(68.35%) | 15<br>(78.95%) |
| 25. For online studies: description of which devices were allowed for participation, e.g., smartphone, PC. | A = .88<br>κ = NA | 4<br>(11.76%) | 6<br>(15.79%) | 2<br>(18.18%) |
| 26. Indication of software used. | A = .65<br>κ = .19 | 18<br>(28.12%) | 19<br>(27.14%) | 7<br>(46.67%) |
| 27. Indication of hardware used. | A = .79<br>κ = .54 | 22<br>(32.35%) | 27<br>(36.99%) | 10<br>(58.82%) |
| 28. Availability of the software for reproducibility. | A = .84<br>κ = .62 | 12<br>(17.14%) | 27<br>(36.49%) | 13<br>(68.42%) |
| 29. Description of characteristics of the experimental setting that, if manipulated, might modulate the size of the effect(s) under investigation (e.g., type of screen, room illumination, distance from the monitor; we will only register whether such characteristics are reported in the paper/supplemental materials; we will not evaluate whether such information should be reported and is missing). | A = .82<br>κ = .57 | 21<br>(30.88%) | 21<br>(31.34%) | 2<br>(11.11%) |
| 30. (only for non-online studies) Specification of whether other people (including the experimenter) were in the room with participants or they were left alone. | A = .74<br>κ = .44 | 18<br>(24.66%) | 16<br>(25.40%) | 4<br>(23.53%) |
| 31. (non-online studies, in case the experimenter stayed in the room) Description of whether the experimenter is a peer or an authoritative person. | A = .88<br>κ = .53 | 10<br>(13.89%) | 13<br>(20.63%) | 3<br>(17.65%) |
| 32. (only for non-online studies, in case other people were present in the room with the participant) Description of what are the roles of the other people in the room. | A = .91<br>κ = .68 | 14<br>(20.00%) | 13<br>(20.97%) | 3<br>(17.65%) |
| 33. Specification of whether there are only one or more sessions. | A = .69<br>κ = .05 | 39<br>(49.37%) | 39<br>(48.75%) | 11<br>(55.00%) |
| 34. (in case there are two or more sessions) Specification of the distance between them. | A = .68<br>κ = -.18 | 19<br>(30.64%) | 15<br>(25.00%) | 3<br>(25.00%) |
| 35. Order of tasks in the experiment. | A = .65<br>κ = .18 | 57<br>(72.15%) | 59<br>(73.75%) | 16<br>(80.00%) |
| 36. (If there are filler tasks) Filler tasks are described in sufficient detail, and there is an indication of their timing. | A = .81<br>κ = .33 | 9<br>(13.23%) | 6<br>(9.37%) | 4<br>(23.53%) |
| **The tasks/the materials** | **ICC = .47** | | | |
| 37. Tasks are described in sufficient detail to be reproduced. | A = .62<br>κ = -.12 | 60<br>(75.00%) | 67<br>(83.75%) | 17<br>(85.00%) |
| 38. Exact instructions are provided. | A = .68<br>κ = .16 | 21<br>(26.25%) | 27<br>(33.75%) | 11<br>(55.00%) |
| 39. It is specified whether instructions are provided in written form or orally. | A = .68<br>κ = .35 | 36<br>(45.00%) | 40<br>(50.00%) | 12<br>(60.00%) |
| 40. The stimuli are provided in sufficient detail (e.g., size, color, sound intensity in dB…). | A = .58<br>κ = .07 | 49<br>(63.64%) | 51<br>(63.75%) | 14<br>(70.00%) |
| 41. The stimuli available, e.g., in the supplemental materials, in the manuscript, or on OSF. | A = .68<br>κ = .37 | 20<br>(25.32%) | 40<br>(50.00%) | 15<br>(75.00%) |
| 42. The exact description of the number of trials, blocks, stimuli per block, number and length of breaks is provided. | A = .72<br>κ = .09 | 63<br>(79.75%) | 58<br>(72.50%) | 15<br>(75.00%) |
| 43. The trial timeline is provided, e.g., inter-stimulus interval; duration of black screen, stimulus duration. | A = .67<br>κ = .33 | 40<br>(52.63%) | 35<br>(44.34%) | 12<br>(60.00%) |
| 44. If there is a practice, it is described in sufficient detail. | A = .82<br>κ = .60 | 21<br>(31.34%) | 28<br>(42.42%) | 5<br>(29.412%) |

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

| | | | | |
|---|---|---|---|---|
| 45. The order of trials is described, e.g., random, fixed for all participants (in this case, is the order provided?). | A = .64<br>κ = .20 | 50<br>(63.29%) | 60<br>(75.95%) | 16<br>(80.00%) |
| 46. It is specified how the responses are given, e.g., orally, keypress, mouse, touchscreen, clicking, swiping, joystick, dynamometer, etc. | A = .68<br>κ = .18 | 57<br>(72.15%) | 55<br>(69.62%) | 16<br>(80.00%) |
| 47. The total duration of the experiment is reported. | A = .80<br>κ = .31 | 22<br>(27.50%) | 20<br>(25.00%) | 2<br>(10.00%) |
| 48. Availability of the tests/questionnaires - we will record whether *all* of the tests and questionnaires are available open source. | A = .72<br>κ = .03 | 42<br>(60.87%) | 49<br>(69.01%) | 13<br>(81.25%) |

*Notes.* N (%). Collabra is not considered in the division by year, as this journal did not exist in 2011. The proportions are considered based on the 'valid cases', or in other words, they exclude the studies for which a certain dimension was irrelevant (e.g., if the paper states that there was no practice, we cannot evaluate the relevant item 48). For reliability coefficients: A = Agreement; κ = Cohen's Kappa; ICC = Intraclass Correlation Coefficient. For item 25 κ is Not Available (NA) because some values are equal to zero.

each paper assigned and the supplemental materials when available, and coded them based on the variables listed in Table 1. After raters completed individual coding, their codes were compared. Areas of disagreement, where two raters had assigned a paper a different evaluation for an item, were discussed among the group of five raters, who reached consensus on the appropriate evaluation. The process of discussion helped to clarify the meaning of the items of the TOM checklist and to generate principles which could be used in future.

For each paper, we computed the TOM score as the ratio between the number of reported features (i.e., those categorized as "1" = present) and the number of relevant features (i.e., those categorized as "1" = present or "0" = absent, i.e., all items except the irrelevant ones). This score indicates the proportion of transparently reported features and ranges from 0 (absence of methods transparency) to 1 (total transparency). Higher values will indicate higher transparency in methods reporting.

### *Hypotheses*

We expected that the TOM score would be higher in 2021 than in 2011 (preregistered). We will also compare the journals to see whether differences in TOM reporting emerge, with higher scores for Collabra (not preregistered).

## Deviations from Preregistration

Compared to the preregistration, we made the following changes: Firstly, we opted to use alternative metrics to assess inter-rater reliability. We preregistered to assess reliability using the Percentage Agreement and the Phi coefficient. As an alternative to Phi, we calculated the intraclass correlation (ICC) of the overall TOM score with the ICC 2k (McGraw & Wong, 1996) and the Cohen's Kappa coefficient (κ; Cohen, 1960). This deviation was motivated by the fact that studies evaluations were performed by five randomly selected raters. We have included results for Agreement, ICC, and Cohen's κ in the relevant tables. The Phi coefficients are also available on OSF for reference. Secondly, to provide a more robust estimation of central tendency, we reported TOM median values along with mean values.

## Results

All the analyses were conducted in R Studio (Posit team, 2023; version 2023.6.2.561).

Overall, the raters reported substantial agreement in their estimate that the time required for evaluating one scientific report, on average, was approximately 30 minutes. The times varied from one report to another, and all evaluators observed a learning effect, meaning that the initial articles assessed took more time than the subsequent ones.

Table S1 presents the counts and percentages of papers that explicitly address each dimension of the checklist for each journal, broken down by year.

## Overall Methods Transparency (TOM score)

For each paper, we computed the TOM score as the ratio between the number of reported features (i.e., those categorized as "1") and the number of relevant features (i.e., those not categorized as "9"). This summarizes the proportion of transparently reported features, and ranges from 0 (absence of methods transparency) to 1 (total transparency), and higher values will indicate higher methods transparency of the paper.[3]

The total TOM score showed moderate reliability (ICC = .57), which we deemed sufficient for the purpose of provid-

3 We report the descriptive statistics for each item and the overall TOM, to provide a faithful summary of the observed data without making unsupported generalizations about the population at large. Due to the non-representative nature of our sample, relying on inferential statistics could lead to unreliable or misleading results, as inferential statistics are designed to draw broader conclusions about a population, assuming the sample is representative. However, the interested reader can find in the Supplementary Materials the results of a binomial generalized linear mixed effects model with publication year as independent variable and TOM scores as dependent variable (Table S2). The results indicate a significant increase in the TOM scores from 2011 to 2021 of a medium-to-large effect size (OR = 3.82, $p < .001$).



Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

**Table 2. Mean, Standard Deviation and Median for TOM for 2011 and 2021 depending on the journal.**

| | | Publication Year | |
|---|---|---|---|
| **Journal** | **Statistic** | **2011** | **2021** |
| JPSP | M (SD) | .41 (.17) | .49 (.14) |
| | Mdn | .44 | .52 |
| EJSP | M (SD) | .34 (.14) | .52 (.18) |
| | Mdn | .34 | .56 |
| Cognition | M (SD) | .45 (.17) | .49 (.16) |
| | Mdn | .48 | .52 |
| JEPG | M (SD) | .36 (.14) | .47 (.14) |
| | Mdn | .34 | .49 |
| Collabra | M (SD) | NA | .55 (.15) |
| | Mdn | NA | .54 |

*Notes*. NA = Not Available (Collabra was not published in 2011).

ing a snapshot regarding the level of methods transparency and evaluating the presence of improvements from 2011 and 2021 (Koo & Li, 2016).

For the four journals JPSP, EJSP, Cognition, and JEPG, the average TOM score of the 80 papers published in 2011 was M = .39 (SD = .16). For the 80 papers published in 2021, it was .49 (SD = .16). The median was Mdn = .39 (Min = .10; Max = .74) in 2011 and Mdn = .52 (Min = .14; Max = .79) in 2021. This shows, in descriptive terms, an improvement in the level of transparency of the scientific reporting in the selected journals. The average TOM score for Collabra was M = .55 (SD = .15), Mdn = .54 (Min = .33; Max = 1.00) and was higher than the TOM scores for the other four journals in 2021.

Table 2 shows the descriptive statistics for TOM, depending on the journal and the year of publication: Papers in all the journals considered have increased the average methods' transparency. Collabra shows the highest average TOM, consistent with the mission of the journal and its publisher, the Society for the Improvement of Psychological Science. To better grasp the variability of TOM scores across papers within journals and year of publication, see Figure 1. It presents a combined dot and violin plot showing the distribution of TOM scores across the five journals. Each journal's scores are compared between the years 2011 (red) and 2021 (blue). For each journal, individual data points represent the TOM scores of respective articles, with the density plots providing an estimation of the distribution shape. The central line in each violin plot indicates the median TOM score for the respective year. It is encouraging to observe a positive trend in the transparency of methods from 2011 to 2021 in all journals. This underscores a shift in the emphasis on methodological transparency in research articles over the decade, with varying degrees of alignment towards the transparency standards set by the flagship journal in this field.

Further, following the preregistration, Table 3 presents the descriptive statistics for TOM depending on the year of publication and psychological field (i.e., cognitive and social psychology).

## Exploring the Nuances: Percentages of Reporting for the Checklist Items

We then evaluated inter-rater reliability for each section of the checklist with ICC 2k, in the same way as for the overall TOM score. The section specific ICCs were moderate for the Participants, Procedure and General Requirements sections, and poor for Task and Materials, and Design section (Koo & Li, 2016). These reliability levels can be interpreted as an indication of the difficulty involved in evaluating the transparency of the methods in a scientific paper. In particular, we observe that the highest level of agreement among evaluators is in the 'Participants' and 'Procedure' sections, which suggests that the degree of transparency in these sections is relatively easier to assess. On the other hand, it appears to be more challenging to determine whether there is sufficient information in the 'Design' section, as the level of agreement among evaluators is the lowest.

The inter-rater reliability for each feature was evaluated using Phi, the Agreement, and Cohen's κ. The results are reported in Table 1. Interrater reliability for the specific items ranged between poor negative values (κ =-.18, item 34) to substantial (κ = .76, item 4) (Landis & Koch, 1977). Specifically, 3 of the 48 values are poor, 16 are slight, 13 are fair, 12 are moderate, and 4 are substantial. While the overall TOM score for an article is a reliable indicator, the reliability of judgments on specific aspects varies a lot from one aspect to another, highlighting the need for caution in drawing conclusions regarding these specific aspects. However, we believe that the data presented in Table 1, which show the proportion of studies reporting the relevant information for each of the dimensions of the methods we evaluated, provide an intriguing overview of the various aspects of methodological transparency. Complying with the preregistration, Table S3 in the Supplementary Materials on OSF shows the proportion of studies reporting relevant information for each dimension by field of study (i.e., social and cognitive psychology).

Regarding Section A (General Minimal Requirements), Table 1 shows a lot of space for improvement. We could

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

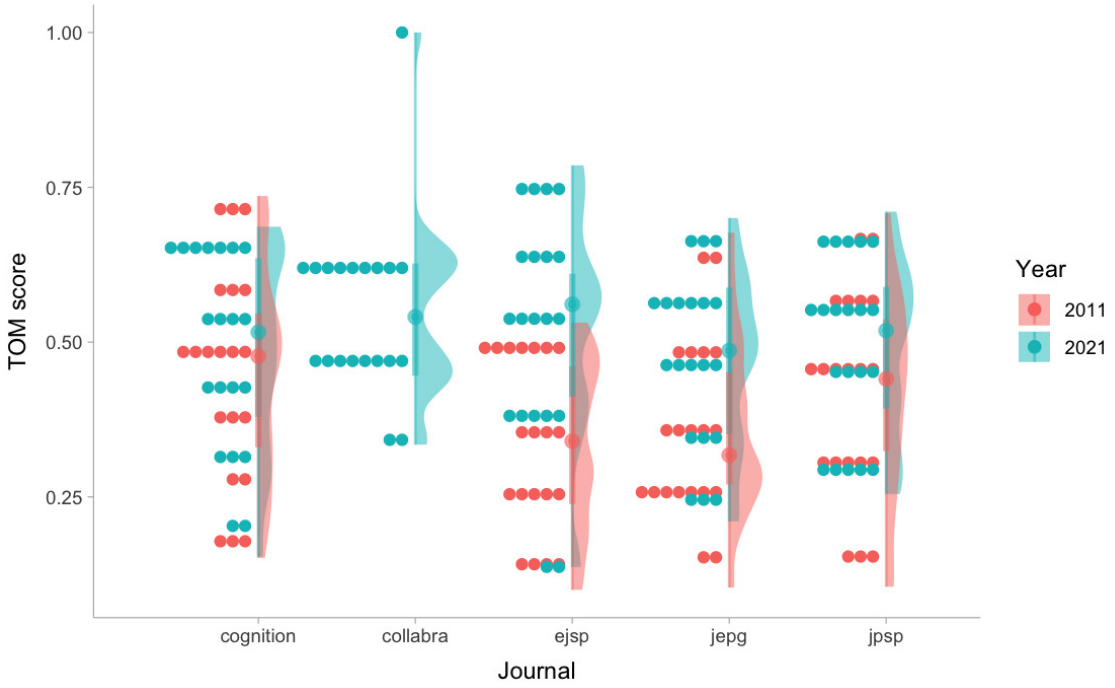



**Figure 1. TOM scores distribution across the different journals and years.**

**Table 3. Mean, Median and Standard Deviation for TOM for 2011 and 2021 depending on the field**

| | Social Psychology | | Cognitive Psychology | |
|---|---|---|---|---|
| | 2011 (n = 40) | 2021 (n = 50) | 2011 (n = 40) | 2021 (n = 50) |
| Mean (SD) | .38 (.15) | .52 (.17) | .41 (.16) | .49 (.14) |
| Median | .40 | .55 | .39 | .50 |

track down statements that all conditions and all dependent variables were reported only in a small proportion of the manuscripts. There has been an important increase in the presence of Open Materials with relevant information for the Methods between 2011 and 2021; however, in 2021 the percentage of papers with Open materials was, on average, 56.3% for the four journals JPSP, EJSP, Cognition, and JEPG, while it was 85% for Collabra: this suggests that these numbers can increase. problematic issue concerns the availability of materials "upon request." It would always be preferable to avoid "on request" availability for materials. People change email addresses or retire, computers break down, etc. Therefore, as time passes, materials available "upon request" become increasingly less available and it is preferable to deposit study materials in open archives with a guarantee of continuity (see Tedersoo et al., 2021). However, even when there are valid reasons to make data available "upon request," the process and the recipients of such requests should be clearly specified. As Table 1 (items 4 and 5) demonstrates, this crucial information was often missing in the papers we analyzed, except for the articles in Collabra, which frequently provided these details (although items 4 and 5 are calculated based on a limited number of Collabra articles since the majority make information available as open materials, as indicated by item 3).

For section B (Participants), Table 1 shows some progress from 2011 to 2021, but several areas for improvement remain. While most papers provide information about the type of population (item 6), age, and gender of participants (item 11), many papers lack other essential details, such as educational level (item 13), which is highly relevant for the suitability of materials for potential reuse and for many phenomena of psychological interest. Additionally, crucial information about the recruitment process (item 8 and item 15) is often missing, which is vital in defining the characteristics of the sample on which the study was conducted.

Section C (Design) is characterized by some elements showing high presence percentages, namely the distinction between manipulated and measured variables (item 17), operationalization procedures (item 18), and the design being either within or between participants (item 19), which is reassuring. However, other crucial elements for interpreting results and potential replications, even in 2021, are only present in about half of the studies (material counterbalancing, item 20; description of control conditions, item 21; and description of manipulation checks, item 22). Includ-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

ing these important elements would further enhance the methodological robustness and reproducibility of research findings.

Moving on to Section D (Procedure), besides the order of tasks (item 35) and information regarding whether the study took place in a laboratory or online (item 24), present in the majority of the examined studies, many other crucial details are found in less than half of them. For instance, despite several empirical pieces of evidence showing that social environment characteristics can influence people's behavior (Steinborn & Huestegge, 2020), information as whether participants were alone or with their peers during the experiment, or whether the experimenter was present or not, is lacking in the majority of the studies we have reviewed, and the data we collected does not indicate a clear improvement in this aspect from 2011 to 2021.

In the last section, which pertains to the task and materials, we observe that a high percentage of studies provide sufficient detail about the tasks (items 37, 42, 45, 46), and the majority make the tests/questionnaires used available or provide bibliographic references to access these materials (item 48). However, other details are only reported in a minority of studies, such as the exact instructions given to participants. A comparison between the data from 2011 and 2021 shows a general improvement, but also in this section, there is room for further enhancement, as indicated by the percentages from Collabra, which are higher for almost all the examined dimensions compared to the overall percentages of other journals. The information on each dimension, separated by journal and by year, is reported on OSF.

## The improved checklist

Given that interrater reliability scores showed the presence of space for improvement, we created an improved checklist. The five co-authors who had initially coded the papers, therefore, discussed the elements that posed the greatest challenges, analyzed potential ambiguities and differing interpretations that emerged only after the checklist had been used to code the papers of this research, and revised the list, modifying the language and adding clarifications and examples. The improved checklist was subjected to evaluation by two raters who were not members of the initial coding team (EC and GF). These two raters read the improved checklist, tested it on three different articles and discussed it with the team. Based on their feedback, further improvements were incorporated. The new, improved list refers to the same dimensions as the previous one, but the items are now expressed in a clearer and less ambiguous language, accompanied by various examples to elucidate the meaning of each dimension.

Specifically, four items were modified to accommodate the evaluation of correlational studies (i.e. studies where there is no experimental manipulation of independent variables or arbitrary assignment to conditions). Twelve items were rephrased to be more precise. Finally, we added new examples to four items: b.12, c.4, d.3 and d.11. All the changes are detailed in Table S4 of the Supplementary Materials.

## Reliability of the Improved Checklist

Table 4 reports the revised checklist, which underwent a check of interrater reliability, where the two new coders independently evaluated the same 40 scientific articles that underwent the initial double-coding using the new list. Raters estimated that the time required to evaluate the manuscripts with the improved checklist was, on average, approximately 30 minutes per paper. As in the original checklist, the time of coding was longer for the initial articles assessed compared to the following ones.

To assess the interrater reliability of the new checklist, we followed the same procedure used in the previous phase. Since the checklist was evaluated by two raters, we opted for the ICC3 (McGraw & Wong, 1996). The ICC for the overall TOM score was .49. This can be considered from poor to moderate (Koo & Li, 2016) and is lower than the value we observed in the evaluation made by the first group of raters. This is an important piece of information, because the key difference between the initial and second group of reviewers is that the first group initially engaged in a comprehensive discussion of each item on the checklist, resulting in a deep understanding of it. The second group of reviewers received an improved checklist with refined terminology and enriched examples, but they did not study it to the same level of detail. Therefore, we interpret this decline in interrater reliability from the first to the second evaluation as an indication that the checklist alone, despite being supplemented with examples and written explanations, is not sufficient for a reliable assessment of methods transparency. In other words, we believe it is crucial to provide the raters with thorough guidance on how to utilize the checklist effectively, to provide them with information on how to use the scale, and to ensure that any uncertainties about its use are addressed before the evaluation begins.

We tested this interpretation of the result by conducting an additional assessment of the reliability of the new checklist. The two raters who conducted the second checklist (EC and GF) reflected retrospectively on what were the major points with which they had difficulties. Based on this reflection, we improved instruction and examples of the TOM checklist, and expanded the stage of coders' familiarization into the following three steps. First, the two new coders who conducted the additional assessment (SP and MS) discussed the checklist item by item with one member of the initial team (FC). Second, they worked jointly with the same member of the initial team to the rating of a familiarization paper, to get a first "hands-on" experience with the checklist, so that any doubt could be addressed in real time. Third, they individually coded two more familiarization papers and afterwards discussed any disagreements and solved them. At this point, the most common source of disagreement was one coder finding a piece of information that had escaped the other, and the second source of disagreement was different interpretations of the more open-ended items (e.g., "Are other relevant demographic characteristics reported in the paper?", where it is up to the coder what counts as "relevant"). The familiarization articles were different from those used for assessing the in-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

**Table 4. Features evaluated in the TOM revised checklist, together with agreement, κ and ICC scores.**

| | 2nd IRR Eval. | | 3rd IRR Eval. | |
|---|---|---|---|---|
| **Checklist items** | **Agreement** | **Cohen's κ** | **Agreement** | **Cohen's κ** |
| **A. General (minimal) requirements** | ICC = .83 | | ICC = .77 | |
| 1. Is it stated that all measured variables and (if applicable) experimental conditions were reported? | .63 | .07 | .72 | .38 |
| 2. Is it stated that materials (i.e. data OR additional analyses OR any other supplemental information) are openly available on a repository such as OSF, Zenodo, etc? | .95 | .89 | .95 | .90 |
| 3. Does the link to materials work? The link is active. materials are present, no permission to access is required. | .95 | .77 | .95 | .77 |
| 4. Are the precise conditions requestors must meet to obtain the materials indicated / is it made clear that the materials will be released unconditionally to all who ask? | 1.00 | 1.00 | 1.00 | 1.00 |
| 5. Is it clear to whom requests for materials are to be directed? | .79 | NA | .81 | .56 |
| **B. Participants** | ICC = .66 | | ICC = .81 | |
| 6. Are the characteristics of the population and all related inclusion or exclusion criteria reported in the paper? | .68 | .33 | .72 | .42 |
| 7. Is the sampling method clearly reported in the paper (e.g., convenience sampling, random sampling)? | .65 | .04 | .85 | .60 |
| 8. Is the recruitment strategy clearly reported in the paper (e.g., University pool, leaflets, post on social media, direct contacts)? | .80 | .59 | .75 | .51 |
| 9. Is it clearly stated whether and how participants were compensated for their participation (e.g., no compensation, course credit, money)? | .88 | .71 | .95 | .89 |
| 10. Is it clearly stated how participants were assigned to study conditions (e.g., random assignment, within participants design)? | .88 | .36 | .94 | .47 |
| 11. Are gender descriptives reported? | .93 | .63 | .98 | .84 |
| 12. Are age descriptives reported? | .90 | .69 | .95 | .84 |
| 13. Are ethnicity descriptives reported? | .90 | .76 | .92 | .83 |
| 14. Are education level descriptives reported? | .82 | .63 | .89 | .65 |
| 15. Are other relevant demographic characteristics reported in the paper? | .79 | .45 | 1.00 | 1.00 |
| 16. Is it clearly indicated how many participants were included in the analysis? | .73 | .49 | .9 | .8 |
| 17. Is it clearly indicated whether any blinding was used (e.g., single blind, double blind)? | .89 | .51 | 1.00 | 1.00 |
| **C. Design** | ICC = .43 | | ICC = .64 | |
| 18. Is it clearly indicated what are the manipulated and measured variables? | .83 | NA | .85 | -.04 |
| 19. Is it clearly indicated how variables were operationalized, i.e., the specific procedures used to measure the target | .55 | .01 | .85 | .35 |

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

| | | | | |
|---|---|---|---|---|
| concepts? | | | | |
| 20. Is the study design clearly described in the paper (e.g., between, within, mixed groups designs)? | .65 | -.05 | .95 | .48 |
| 21. Is it clearly reported in the paper what was counterbalanced between participants (e.g., no counterbalancing, settings, tasks, stimuli)? | .78 | .57 | .75 | .52 |
| 22. Is it clearly indicated what participants did in the control condition? | 1 | NA | NA | NA |
| 23. Is it clearly indicated whether and what manipulation check was used (e.g., no manipulation check, follow-up questions) ? | .86 | .67 | .97 | .92 |
| **D. Procedure** | ICC = .28 | | ICC = .77 | |
| 24. Is it reported what information was provided at invitation to participate? | .63 | .18 | .88 | .71 |
| 25. Is it clearly reported where the study took place (e.g., in the lab, online survey)? | .55 | .18 | .98 | .95 |
| 26. Is it clearly indicated which devices were used to collect data (e.g., computer, tablet, smartphone, pen and paper)? | .65 | .17 | .98 | .94 |
| 27. Is it clearly indicated what hardware was used (e.g., monitor size and refresh rate, eye tracker sampling rate)? | .82 | .65 | 1.00 | 1.00 |
| 28. Is it clearly indicated what experimental software was used (e.g., Inquisit, Matlab, Qualtrics, tailored apps)? | .72 | .44 | .88 | .76 |
| 29. Is the experimental script available for reproducibility? | .76 | .29 | .83 | .47 |
| 30. Are the characteristics of the experimental setting reported? | .71 | .18 | .80 | .62 |
| 31. Is it clearly reported whether there were other people in the room with the participants (e.g., researcher, confederate)? | .80 | .58 | .91 | .68 |
| 32. Is it clearly indicated whether the experimenter was a peer or an authoritative person (e.g., professor, post-graduate, physician)? | .90 | .52 | .88 | .30 |
| 33. If other people besides the experimenter were present in the room, is it clearly stated what their roles were? | .97 | .84 | .97 | .87 |
| 34. Is the duration of the study session(s) clearly indicated? | .75 | .31 | .90 | .65 |
| 35. Is it clear from the description of the procedure whether there were only one or several sessions? | .68 | .05 | .75 | .13 |
| 36. Is it clearly indicated how far apart the experimental sessions were? | .60 | .00 | .60 | .17 |
| 37. Is the order of administration of the tasks/instruments (e.g., manipulation, measures) clearly indicated? | .45 | .05 | .95 | .64 |
| 38. Are filler tasks clearly described? | .67 | NA | NA | NA |
| **E. Task** | ICC = .33 | | ICC = .76 | |
| 39. Are the exact (verbatim) instructions reported for all tasks? | .76 | .44 | .47 | .08 |
| 40. Is it clearly indicated how task instructions were provided (e.g., written, orally)? | .62 | .28 | .76 | .50 |

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

| | | | | |
|---|---|---|---|---|
| 41. Are the stimuli described in detail (e.g., size, color, sound intensity in dB, position on the screen)? | .52 | .17 | .95 | .90 |
| 42. Are the exact stimuli openly available? | .65 | .35 | .88 | .75 |
| 43. Are the task characteristics clearly reported in the paper (e.g., number of trials and blocks, number of stimuli per block, number and length of breaks)? | .77 | -.06 | NA | NA |
| 44. Is the trial timeline specified in the paper (e.g., inter-stimulus interval; duration of black screen, stimulus duration)? | .67 | .31 | .93 | .63 |
| 45. Is the order of the trials in each task clearly specified (e.g., random, fixed for all participants)? | .68 | .28 | .95 | .83 |
| 46. Is it clearly specified how participants provided their answers (e.g., orally, keypress, mouse, touchscreen, clicking, swiping, joystick, dynamometer)? | .74 | .37 | .88 | .72 |
| 47. Is it clearly specified whether feedback was provided after participants answer? | .86 | .68 | .95 | .89 |
| 48. Is it clearly specified what participants had to do in the task practice phase? | .75 | .40 | .82 | .39 |
| **F. Open Materials** | ICC = .36 | | ICC = .54 | |
| 49. Are the tests/questionnaires used in the study available? | .80 | .55 | .83 | .57 |
| 50. If the study includes questionnaires: is there a bibliographic reference for all the questionnaire? | .85 | .41 | NA | NA |
| 51. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Are the exact instructions reported for the questionnaires? | .75 | .50 | 1.00 | 1.00 |
| 52. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Is it clearly indicated how task questionnaire instructions were provided (e.g., written, orally)? | .44 | .29 | NA | .00 |
| 53. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Is it clearly specified how participants provided their answers (e.g., written, orally)? | .40 | .15 | NA | .00 |
| 54. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Are all the items reported? | .89 | .61 | 1.00 | 1.00 |
| 55. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Is it clearly indicated what was the order of the items? | .33 | NA | .50 | -.29 |
| 56. For questionnaires without reference (answer to previous questionnaire.reference is 0 or 9): Is it clearly indicated what were the response options to the items? | .63 | NA | NA | .00 |
| 57. If the study includes observation | NA | NA | NA | NA |

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

| | | | | |
|---|---|---|---|---|
| grids (e.g. those used to assess child behavior in a given situation): is there a bibliographic reference for the observation grid? | | | | |
| 58. For observation grids without reference (answer to previous grid.reference is 0 or 9): Are the exact instructions for observation reported? | NA | NA | NA | NA |
| 59. For observation grids without reference (answer to previous grid.reference is 0 or 9): Are all the areas of observation described? | NA | NA | NA | NA |
| 60. For observation grids without reference (answer to previous grid.reference is 0 or 9): Are the response options for each dimension clearly indicated? | NA | NA | NA | NA |

Notes: IRR= Interrater Reliability; NA = Not Available. The Agreement and k coefficients are NA for certain items because some values are equal to zero.

terrater reliability. However, they were chosen using similar criteria as the target articles: three of the five target journals were randomly selected and, from each journal, the most recent (at the time of familiarization) standard research paper was used for new coder training. Their DOIs are indicated in the list of included papers (Supplementary Material). After this familiarization, the two new coders expressed confidence with the process and moved on to coding the target articles for the new estimate of interrater reliability.

Supporting our reasoning that thorough guidance could lead to increased reliability of the scale, this third evaluation evidenced a sensible increase in interrater reliability for the overall TOM score, which increased to ICC = .84, which is considered an indication of good reliability according to Koo & Li (2016).

Table 4 reports the Agreement and Cohen's κ scores for each feature, and ICC for each section, for the second and third check of the reliability.

Considering in specific the third evaluation of interrater reliability, besides the good reliability for the overall score, it is also interesting to note that the ICC is at least moderate for all sections. The interrater reliability for the individual features ranged between poor values (K = -.29, item 55) to perfect values (K = 1). These values also depend on the number of papers specifically assessed for each question (available on OSF). In some cases with a low number of papers considered, extreme negative values occurred. Overall, 2 out of the 49 dimensions showed poor reliability, 3 slight, 4 fair, 10 moderate, 12 substantial, and 18 excellent reliability.

## Discussion

The goal of this work was twofold: to assess the feasibility of investigating the transparency of the reporting of the methods in experimental research in psychology using the checklist and TOM score, and to provide insight into the methods transparency in articles published over a ten-year period in a selection of journals.

As concerns the feasibility of the method, the results are encouraging. Both with the initial and the improved checklist, evaluators agreed that the time required to assess a scientific experiment report was approximately half an hour. This indicates that this research method requires a time investment, but it can be manageable, especially if the research is done through a collective effort. It also suggests that an important future step could be to streamline the checklist by distinguishing essential from eliminable items. Both when the initial evaluators used the original checklist, following discussion among themselves, and when the new evaluators employed the enhanced checklist, especially after a thorough familiarization phase, the reliability of the overall TOM score, as well as a significant portion of the individual items, was satisfactory. This corroborates the feasibility of a collaborative research effort on investigations regarding method transparency.

We also collected proofs of construct validity: Firstly, the TOM index increased from 2011 to 2021, a trend expected given the heightened awareness and reflection psychology has undergone over the past decade. Moreover, the journal from which the highest level of transparency commitment would be anticipated, namely Collabra, is precisely the one where articles achieve, on average, the highest scores.

With regard to the objective of gathering insights on the level of transparency in the methods of research articles published between 2011 and 2021, our examination of a selection of top-tier journals reveals that, despite the fact that an increase in methods transparency emerged, there still is much room for improvement, as the average and median TOM scores indicate that typically, only half of the information is reported in 2021.

An inspection of the items' values also shows the areas where there is most space for improvement. Some of the less often reported aspects are relatively easy to incorporate. For instance, one of these is the statement that all conditions and all dependent variables are reported. This addition would simply involve an extra sentence in the article and would offer readers and reviewers greater assur-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

ance against practices like cherry-picking results. To give an example of another area of improvement, results indicate that in 2021 much information on the social setting of the experiment was still missing: Often it was not sufficiently clear from the report whether the participants in laboratory settings performed the experiment alone or in the presence of other people, and what were the roles of the other people present in the setting. However, the presence of other individuals importantly affects the way humans think and behave (Klein et al., 2012; Van Bavel et al., 2016).

Generally, it is important to stress that seemingly minor protocol deviations in an experiment can substantially change the results. For instance, Chen and colleagues (2021) found that receiving the instructions only written on a computer screen, or read aloud to them as well, substantially impacted the results of their study. When people participate in psychology experiments, they actively engage in the process of constructing meaning, interpreting, and shaping their experience within the experimental context. They form beliefs that can influence how they react to the situation and, in the end, the research results. This construction of meaning can be influenced by various cues provided during the study. These cues include how the study is presented upon invitation, the precise instructions they receive, the filler tasks they answer to, and more. However, less than one-third of the scientific reports we examined indicated how the study was presented to participants, and an even lower percentage reported what were the filler tasks used.

In general, human behavior is susceptible to many variables, at times seemingly inconsequential (Kahneman, 2014), so while the optimal methods section should include all the information necessary for a direct replication and allow to understand how the research was conducted, deciding what details are pertinent is not as straightforward as it may initially appear (see, e.g., Arehalli & Wittenberg, 2021, for an example of effects of filler tasks in psycholinguistics). All the more reason why it's important to encourage authors to be transparent in describing the context of their studies and sharing their materials as much as possible.

## Limits and Opportunities for Future Research

The sample of papers that we investigated sets a boundary for the types of inferences we can make. We considered only five top-tier journals in certain areas of psychology. If there is such a wide scope for improvement in articles published in these journals, analogous space for improvement is most probably present also in papers published in other journals. However, conducting further research with a larger sample of publications would be crucial. It would also be highly interesting to include articles describing different types of studies beyond experimental research. With a more extensive and representative corpus of articles, comparisons could shed light on which areas of psychology exhibit greater methodological transparency and which may require targeted efforts to promote a culture of transparency. Similarly, with a broader dataset, it would be possible to investigate whether there are author characteristics (e.g., academic seniority, geographic origin, or others) associated with higher levels of methods transparency. In essence, we believe that this work can be considered a preliminary step, a method test, and that much work lies ahead.

## Improving the TOM checklist: learning from our experience

Refining and using the checklist has required us to face its limitations. As mentioned when presenting the results of the second and third rounds of evaluation, the checklist has proven a reliable tool, but there is still room to improve its usability. Perhaps the most critical issue was minimizing the degree of subjectivity involved in the rating process. In this work we described the development of an improved version of the checklist, with items reworded to reduce ambiguity and the addition of example cases, and we believe there are benefits to be gained from continuing on this path. For instance, some items could still be refined to minimize the need for interpretation. To provide an example, item 6: "Are the characteristics of the population and all related inclusion or exclusion criteria reported in the paper?" requires that the rater makes a subjective decision on whether those reported are all the inclusion and exclusion criteria applied. An item addressing the inclusion and exclusion criteria with a lower degree of subjectivity could be: "Does the report clearly indicate the inclusion and exclusion criteria?".

While we are cautious about further altering item contents, we intend to incorporate the feedback received from the second and third group of raters as well as from reviewers, refining the language and providing accompanying information that can help to minimize raters' uncertainty. As a first step, we have created a document, to be paired with the checklist, that contains instructions for use and answers to frequently asked questions. We intend this to be a "living document", regularly updated to offer guidance on issues and special cases that checklist users may encounter in their evaluation work. The current version of the document is available on the OSF project page.

## How to Enhance Methodological Transparency in Reporting: Potential Interventions

If we want authors to be more transparent in reporting the methods used in their studies, a diversified approach is essential, one that operates on multiple fronts and includes interventions acknowledging that different researchers, disciplinary and sub-disciplinary fields, and journal readerships may be at different stages in the reflection and on the issue.

*Firstly, it is crucial to cultivate a culture of transparency.* This implies increasing researchers' awareness of the benefits that transparency practices can have for science. The process can also be facilitated if good practices in sharing and transparency are promoted as virtuous behaviors that are being increasingly adopted by the scientific community. However, this shift in culture and behavior is more likely to happen if prescriptive norms are enforced by publishers and funders. Such norms would require methodological trans-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

parency as a prerequisite for publication or funding unless valid reasons exist for non-disclosure. Previous research has evidenced an association between journal policies on data-sharing and increases in such practice (Hardwicke et al., 2018; Kidwell et al., 2016; Naudet et al., 2018; Nuijten et al., 2017; Rowhani-Farid & Barnett, 2016). Although we are not aware of similar research on methods transparency, also in this realm an analogous relation is plausible.

*What are journals' and funders' policies regarding Open Materials?* One way to evaluate journals' and funders' policies concerning open science practices is through their level of compliance with the Transparency and Openness Promotion guidelines (TOP; Nosek et al., 2015). The TOP guidelines are shared standards for open practices across journals; they describe which open science practices journals require from the authors who aim to publish (Nosek et al., 2015). These guidelines address various aspects of open practices, besides Research Materials Transparency and Design and Analysis Transparency, which are particularly relevant to the purposes of this work. For each dimension, they differentiate three levels of journal commitment: Level 1 consists in offering incentives for adopting such practice, level 2 indicates a stronger expectation for authors, and at level 3 the journal enforces the practices. As of May 19th, 2024, 3080 journals have been evaluated according to their degree of adherence to the TOP guidelines (https://www.topfactor.org/). As concerns the dimension of Research Materials Transparency, a score of 2 is assigned when the journal requires authors to post materials to a trusted repository and identify any exceptions at article submission, and a score of 3 when it requires not only posting the materials but also that the analysis be reproduced independently prior to publication. Out of the 3080 journals, only 111 (3.60%) have received a score of 2 or 3 on this dimension, and only an additional 460 (14.90%) require that the article states whether materials are available and, if so, where to access them. Regarding the other TOP guideline relevant to the transparency of methods, namely Design and Analysis Transparency, a score of 2 is granted when the "Journal requires adherence to design transparency standards for review and publication" and 3 when this requirement is enforced. Overall, out of the 3080 evaluated journals, 195 (6.33%) were granted a score of 2 or 3, and another 826 (26.82%) scored 1, meaning that the submission guidelines articulated design transparency standards. As concerns Funders, the OSF site (https://osf.io/kgnva/wiki/Funders/), as of May 19th, 2024, lists 25 public and private funding agencies whose policies include reference to open science practices.

Overall, these numbers are encouraging as they highlight that journals and founders are beginning to place importance on methods transparency. This can be an important driver towards greater transparency, but in fact, much remains left to individual authors. Not only do the majority of funders and journals leave the choice to individual authors, but the policies are often vague as to what aspects of the methods should be disclosed and how.

*In addition to the stick, the carrot.* Another useful measure can be the use of badges. For instance, Kidwell and colleagues (2016) found that when the journal Psychological Science offered authors an open materials badge, there was a subsequent increase in the sharing of materials. Importantly, the works of Hardwicke and colleagues (2021) and Crüwell and colleagues (2023), which concerns in particular open badges for data sharing, shows that badges alone are not sufficient to ensure that the intended consequences in terms of transparency and reproducibility are met. Badges can serve as a transitional tool to incentivize good practices (see Hardwicke & Vazire, 2023). However, it is important that they are accompanied by clear and unambiguous indications on the minimal requirements to obtain the badge, guidance for the authors, and check that the minimal requirements are satisfied, to ensure they do not become purely performative actions.

But, in our opinion, the most important incentive to create a real change in the behavior of authors regards promotion and tenure, which should be better aligned with the quality, instead of quantity, of research (see Moher et al., 2018).

Importantly, the impact of reform initiatives needs to be iteratively evaluated to understand what works to encourage transparency and what does not work. We hope our TOM can be helpful to this end.

*A serendipity note.* Employing the checklist to assess the work of fellow authors has notably heightened our awareness of methodological transparency, and it has profoundly impacted our own approach as authors in composing scientific reports. Consequently, it appears conceivable that, beyond serving as a measurement instrument as in the present work, a checklist during our own writing endeavors or in aiding other authors as editors or reviewers as envisioned in the PECANS initiative (PECANS, 2024), the application of this tool could serve as a potent intervention to cultivate authors' proficiency in producing research works characterized by enhanced transparency.

## Conclusions

Here, we offer a snapshot of the level of methods transparency in the scientific papers published by five top-tier journals in psychology, and an initial check of whether an improvement has occurred compared to ten years ago. Feeling the pulse of scientific transparency practices, and tracking the progress over time, will be useful to gauge how successful the scientific community is in this endeavor, in what aspects we need more effort, and more generally speaking, to provide relevant information for future reflections on the necessary policies. Meta-research is important to evaluate reform initiatives' impact and provide useful data for thinking about the next stages. We provide relevant data not only for such a reflection but also a proof of concept of a methodology for measurement that can be useful for future studies.

## Competing Interests Statement

The authors state that no competing interests exist.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

## Author Contributions

Conceptualization: C.Z., M.V., A.F., F.C., E.R., and F.G.
Data curation: M.V. and A.F.
Formal analysis: C.Z., M.V., and A.F.
Funding acquisition: C.Z.
Investigation: C.Z., M.V., A.F., F.C., E.R., F.G., E.C., G.F., S.P., and M.S.
Methodology: C.Z., M.V., A.F., F.C., E.R., and F.G.
Project administration: C.Z. and F.C.
Software: M.V. and A.F.
Supervision: C.Z., F.C., and E.R.
Validation: E.C., G.F., S.P., and M.S.
Visualization: M.V. and A.F.
Writing - original draft: C.Z., M.V., A.F., and F.C.
Writing - review & editing: C.Z., M.V., A.F., F.C., E.R., F.G., E.C., G.F., S.P., and M.S.

## Data Accessibility Statement

All data, materials, and analysis scripts related to this study are openly available on OSF (https://osf.io/etkh5/).

Submitted: November 06, 2023 PDT, Accepted: June 21, 2024 PDT

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

## References


Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PloS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Arehalli, S., & Wittenberg, E. (2021). Experimental filler design influences error correction rates in a word restoration paradigm. *Linguistics Vanguard*, *7*(1), 20200052. https://doi.org/10.1515/lingvan-2020-0052

Barrios, M., Guilera, G., Nuño, L., & Gómez-Benito, J. (2021). Consensus in the delphi method: What makes a decision change? *Technological Forecasting and Social Change*, *163*, 120484. https://doi.org/10.1016/j.techfore.2020.120484

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Chen, R., Chen, Y., & Riyanto, Y. E. (2021). Best practices in replication: a case study of common information in coordination games. *Experimental Economics*, *24*, 2–30. https://doi.org/10.1007/s10683-020-09658-8

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., 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. https://doi.org/10.1177/09567976221140828

Ejelöv, E., & Luke, T. J. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, *87*, 103937. https://doi.org/10.1016/j.jesp.2019.103937

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: evidence from a study registry. *Soc. Psychol. Personal. Sci.*, *7*, 8–12. https://doi.org/10.1177/1948550615598377

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. https://doi.org/10.1177/1088868313507536

Grahe, J. (2018). Another step towards scientific transparency: Requiring research materials for publication. *The Journal of Social Psychology*, *158*(1), 1–6. https://doi.org/10.1080/00224545.2018.1416272

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: an observational study. *Royal Society Open Science*, *8*(1), 201494. https://doi.org/10.1098/rsos.201494

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hardwicke, T. E., & Vazire, S. (2023). Transparency is now the default at Psychological Science. *Psychological Science*, 09567976231221573. https://doi.org/10.1177/09567976221140828

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *Royal Society Open Science*, *7*(2), 190806. https://doi.org/10.1098/rsos.190806

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. https://doi.org/10.1177/0956797611430953

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology*, *45*, 299–311. https://doi.org/10.1027/1864-9335/a000202

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing*, *53*(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x


Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PloS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low Hopes, High Expectations: Expectancy Effects and the Replicability of Behavioral Experiments. *Perspectives on Psychological Science*, *7*(6), 572–584. https://doi.org/10.1177/1745691612463704

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174. https://doi.org/10.2307/2529310

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. E. (2019). An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. *Perspectives on Behavior Science*, *42*, 33–57. https://doi.org/10.1007/s40614-019-00193-3

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. https://doi.org/10.1177/1745691613491437

Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics and Its Application*, *4*, 109–122. https://doi.org/10.1177/1745691613514755

Markus, H. R., & Stephens, N. M. (2017). Editorial overview: Inequality and social class: The psychological and behavioral consequences of inequality and social class: A theoretical integration. *Current Opinion in Psychology*, *18*, 4–12. https://doi.org/10.1016/j.copsyc.2017.11.001

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLoS Biology*, *16*(3), e2004089. https://doi.org/10.1371/journal.pbio.2004089

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *Bmj*, *360*. https://doi.org/10.1136/bmj.k400

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. S. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*(1), 31. https://doi.org/10.1525/collabra.102

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

PECANS. (2024). *Preferred Evaluation of Cognitive And Neuropsychological Studies - The PECANS statement for human studies*. https://doi.org/10.17605/OSF.IO/JVZE5

Posit team. (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. http://www.posit.co/

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7

Rowhani-Farid, A., & Barnett, A. G. (2016). Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open*, *6*(10), e011784. https://doi.org/10.1136/bmjopen-2016-011784

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Steinborn, M. B., & Huestegge, L. (2020). Socially alerted cognition evoked by a confederate's mere presence: analysis of reaction-time distributions and delta plots. *Psychological Research*, *84*, 1424–1439. https://doi.org/10.1007/s00426-019-01143-z

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology: A systematic review of prevalence estimates and new empirical data. *Zeitschrift für Psychologie*, *227*(1), 53–63. https://doi.org/10.1027/2151-2604/a000356

Suter, W. N. (2020). Questionable Research Practices: How to Recognize and Avoid Them. *Home Health Care Management & Practice*, *32*(4), 183–190. https://doi.org/10.1177/1084822320934468

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Trevelyan, E. G., & Robinson, N. (2015). Delphi methodology in health research: how to do it? *European Journal of Integrative Medicine*, *7*(4), 423–428. https://doi.org/10.1016/j.eujim.2015.07.002

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1), 1. https://doi.org/10.1525/collabra.74

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies. Review and implications for future quality assurance. *Technological Forecasting & Social Change*, *79*, 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025

## Supplementary Materials

## Supplemental Material

Download: https://collabra.scholasticahq.com/article/121243-assessing-the-transparency-of-methods-in-scientific-reporting/attachment/236396.docx?auth_token=RqiGF8n4NRfE_sVqmzvn

Downloaded from http://online.ucpress.edu/collabra/article-pdf/10/1/121243/840834/collabra_2024_10_1_121243.pdf by University of Salento user on 02 June 2025