UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Markant, D. B., & Galati, A. (2023). Development of a Concept Inventory on Open and Transparent Research Practices. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.75226

Methodology and Research Practice

Development of a Concept Inventory on Open and Transparent Research Practices

Douglas B. Markant[1], Alexia Galati[1][a]

[1] Psychological Science, University of North Carolina at Charlotte NC, US

Keywords: open science, research methods, concept inventory, credibility crisis, replication crisis, assessment, pedagogy

https://doi.org/10.1525/collabra.75226

Collabra: Psychology

Vol. 9, Issue 1, 2023

Over the past decade psychology researchers have begun adopting practices that promote openness and transparency. While these practices are increasingly reflected in undergraduate psychology curricula, pedagogical research has not systematically examined whether instruction on open science practices improves students' conceptual understanding of research methods. We developed the Open Science Concept Inventory (OSCI) to evaluate the impact of integrating open science practices into research methods courses. First, we created a set of hypothetical dilemmas related to a range of open science concepts and elicited open-ended responses from undergraduates (N = 64, Study 1). Based on the responses, we created a 40-item multiple-choice questionnaire, which we administered to a new group of participants (N = 262, Study 2) and used item response theory to select 33 items for the final OSCI. Finally, in two implementation rounds across two semesters (Study 3, total N = 37), we evaluated students' learning gains with the OSCI in a pre-test/post-test design. The implementation rounds involved new materials on open science for a psychology research methods course, including video lectures that situated questionable research practices in the current norms of science and introduced emerging solutions. After excluding extremely fast survey responders, an exploratory analysis showed learning gains among students who expended appropriate effort when completing the OSCI. By systematically evaluating a tool that is easily integrated into existing curricula, we aim to facilitate the adoption of open science practices in undergraduate instruction and the assessment of students' conceptual foundations for conducting robust and transparent research.

Introduction

In recent years psychology researchers have begun espousing practices that promote openness and transparency, including sharing data, providing reproducible analysis code, and pre-registering hypotheses and analysis plans. These practices have emerged in response to concerns over failures to replicate previous findings, conventional practices that undermine the robustness of published research, and outright fraud (Nelson et al., 2018). In concert with these widespread methodological shifts, academics have sought to integrate the lessons of the replication crisis and open science movement into undergraduate psychology education (Blincoe & Buchert, 2020; Chopik et al., 2018; Frank & Saxe, 2012; Sarafoglou et al., 2020). Still, the extent to which open science research practices are reflected in the teaching of research methods in psychology remains unclear (Jarke et al., 2022). Moreover, little is known about how introducing undergraduates to open science practices affects their conceptual understanding of quantitative research methods and their broader attitudes toward psychological research.

To bridge this gap, we developed and evaluated a targeted assessment instrument, the Open Science Concept Inventory (OSCI), which can be used to evaluate the impact of teaching open science practices in research methods courses. In what follows, we first present a brief overview of emerging open and transparent research practices, addressing some of the key concepts that can be incorporated in a quantitative research methods curriculum and are assessed by the OSCI. We review some of the pedagogical efforts to date that have sought to integrate open and transparent research practices in undergraduate courses. Then, we describe the steps we took to develop the OSCI to evalu-

[a] The authors contributed equally to this work.

Correspondence concerning this article should be addressed to Doug Markant or Alexia Galati, Department of Psychological Science, University of North Carolina at Charlotte. Email: dmarkant@uncc.edu, agalati@uncc.edu

ate students' conceptual understanding of these topics and to assess the impact of new materials on open science in a research methods course.

The Emergence of Open and Transparent Research Practices and Their Integration in the Classroom

Psychology has undergone a transformation as a field over the past decade. A confluence of events around 2010–2012, including revelations of replication failures and data fraud, forced researchers to reflect on the best practices for collecting, analyzing, and reporting data (Nelson et al., 2018). Concerns during this "replication crisis" centered on a number of questionable research practices (QRPs), including: failing to report all conditions, measures, or experiments in a study (Simmons et al., 2011), "fishing expeditions" during statistical analysis (e.g., conducting multiple tests and reporting only those that reach statistical significance, referred to as *p-hacking*; Simonsohn et al., 2014), reporting unpredicted results as if they had been hypothesized in advance (i.e., "hypothesizing after the results are known" or HARKing; Kerr, 1998), and other practices that exploit researcher degrees of freedom (Wicherts et al., 2016). In contrast to cases of data fabrication or fraud, which appear to be relatively rare (Fanelli, 2009), QRPs had been culturally accepted practices (John et al., 2012) stemming from pervasive incentive structures in science that reward the publication of positive, confirmatory, or novel findings.

In response to the replication crisis and the QRPs fueling it (attested also in other fields, such as biomedical research; Ioannidis, 2005; criminology, Pridemore et al., 2018; ecology and evolutionary biology, Kelly, 2019), researchers started embracing new methodological practices that promote transparency and openness. These practices include voluntarily posting data, sharing analysis code, pre-registering hypotheses, and documenting methods and results fully through open science tools (Gernsbacher, 2018a; Klein et al., 2018).

These practices have also begun to make their way into the psychology undergraduate curriculum. Understanding the replication crisis and open science is especially important for students who will pursue further training in psychology and will benefit from instruction on the practical skills and conceptual foundations for conducting robust research. However, regardless of their eventual career path, this instruction is also essential for all students to make reasoned judgments about the credibility of psychological research and scientific evidence more broadly (Sarafoglou et al., 2020). Pownall and colleagues (2022) review recent approaches for incorporating open and transparent practices in courses and find benefits in terms of practical skills for conducting research and student engagement, as well as some evidence for improved attitudes toward psychological research. For example, in a study by Chopik and colleagues (2018), after a one-hour lecture that covered issues surrounding replicability and open science, undergraduates reported being more skeptical about findings from specific studies but expressed increased confidence in psychology as a field (e.g., considering it more similar to the natural sciences).

Other educators have shifted the focus of student projects to replications of published findings (Frank & Saxe, 2012; Jekel et al., 2020; Wagge et al., 2019). Advocates of this approach argue that didactic replications provide a rich pedagogical context for undergraduates to learn the basic tools of science, while also contributing independent replications of existing findings. More recently there have also been educational efforts focused on large-scale preregistered replications and extensions of classic findings in judgment and decision making and social psychology (Collaborative Open-science Research, 2022). Beyond replications, preregistrations of projects pursuing novel research questions have also been integrated in undergraduate courses; for example, in capstone research courses (Blincoe & Buchert, 2020). These efforts can help the field by making replication mainstream (Gernsbacher, 2018b), demystifying the process of preregistration, and highlighting the benefits of openness and transparency for the next generation of science producers and consumers.

However, in most cases, these educational efforts have not been accompanied by systematic evaluation of whether they in fact improve students' conceptual foundations for conducting research (Pownall et al., 2022; see also Linn et al., 2015). After targeted instruction, researchers do not typically examine directly learning gains about open science concepts. For example, although Chopik et al. (2018) found that students endorsed open research practices after their instructional intervention, it's unclear whether students could reason about the motivating problems with QRPs or apply the concepts from the lecture to new situations. Systematic evaluations are needed to identify approaches for undergraduate research methods instruction that both train practical skills and foster the conceptual understanding necessary to apply those skills in other circumstances outside the classroom (Linn et al., 2015). Toward that end, we developed and evaluated a tool which we then use to systematically assess the effectiveness of a targeted curriculum on open science integrated in a research methods course.

At this juncture, we should acknowledge that the scope and level of detail of a targeted curriculum on open science is difficult to prescribe. Some of the "best practices" emerging from the open science movement have been the subject of debate, highlighting that methodological reform requires nuance and rigor (Devezer et al., 2021). For example, preregistration is typically described as a step in the research pipeline that reduces researchers' degrees of freedom and improves diagnosticity (in discriminating a particular hypothesis from its alternatives) by distinguishing exploratory and confirmatory predictions. However, not all researchers agree about the appropriateness of preregistration. Some argue that preregistration does not, in fact, improve the diagnosticity of statistical tests, because statistical models typically approximate the underlying theory poorly (Szollosi et al., 2020). Others make the point that distinguishing between justified and arbitrary choices of hypotheses and analyses (i.e., between discovery-oriented

and theory-testing research) is more useful than distinguishing between exploratory and confirmatory predictions (Oberauer & Lewandowsky, 2019). Other scholars argue that, although reproducibility is desirable, the relationship between reproducibility and scientific discovery is not well understood (Devezer et al., 2019, 2021). For example, Devezer and colleagues (2019) showed through simulations that the scientific process may not converge on the truth even when results are reproducible. Related perspectives frame the replication crisis as a crisis in theory rather than a crisis in methods. Many researchers argue that scientific reform requires a stronger grounding in theory development, specification of formal models, and transparency in metamethods (Guest & Martin, 2021; Muthukrishna & Henrich, 2019; Smaldino, 2017; Szollosi et al., 2020; van Rooij & Baggio, 2020). Finally, the current scientific reform movement has been criticized for being exclusionary of epistemic traditions rooted in qualitative research (Field et al., 2021). For example, linking replicability with scientific legitimacy and trustworthiness is problematic (Praat et al., 2020; Field et al., 2021), as it makes assumptions about the dominant methods in science (including psychology). It also ignores the means by which qualitative researchers achieve rigor, such as maintaining detailed metadata, having multiple coders (Lorenz & Holland, 2020), member checking, and engaging in reflexivity (Field et al., 2021).

In sum, there are multiple pathways to improve the robustness and rigor of scientific research. A key challenge for instructors of open science (and research methods more generally) is to avoid prescribing a set of one-size-fits-all procedures, and instead support students in learning to reason about the research process and to identify potential limitations and opportunities for bias. In the present work, we focus on a set of widely acknowledged factors that have contributed to the reproducibility crisis and which are appropriate for undergraduate courses on quantitative research methods. However, we recognize that the epistemic roots of psychology and other scientific disciplines are diverse and draw on other methodological traditions such as qualitative and mixed methods approaches.

The Current Study

In this preregistered set of studies,[1] we developed and evaluated the Open Science Concept Inventory (OSCI), which we then used to assess the effectiveness of a targeted curriculum on open science in an undergraduate psychology research methods course. To develop the OSCI, we first created an initial set of vignettes describing dilemmas faced by a researcher or other stakeholder. In Study 1, we administered these vignettes to participants recruited from a psychology subject pool and elicited open-ended responses about how researchers should respond to each scenario. By qualitatively coding participants' elicited responses, we generated multiple choice options which included a "best practice" and common misconceptions for 40 retained vignettes. In Study 2, these multiple-choice items were presented to a new group of participants, whose responses were analyzed with Item Response Theory (IRT) to select the final set of items for the OSCI. Finally, Study 3 was a preliminary "proof of concept" for using the OSCI as an assessment tool: we used a pre-/post-test design across two semesters (Fall 2019 and Spring 2020) to examine performance on the OSCI before and after students completed open science modules in a psychology research methods lab course. We evaluated whether the integration of open science instruction improves students' conceptual understanding of these topics. Following previous studies (Chopik et al., 2018), we also assessed whether students' attitudes toward psychological research and reported self-efficacy would improve by the end of the course. We predicted that students would experience domain-relevant learning gains after completing the curriculum on open and transparent research practices, as assessed by the OSCI. Moreover, we predicted that, by the end of the course, students would report increased self-efficacy in research and improved attitudes toward psychology as a field.

Study 1

Participants

Sixty-four students (38 identified as female, 26 as male; age M = 21.4 years, SD = 3.8) from UNC Charlotte in the United States completed the survey.[2] We did not collect information about the participants' race and ethnicity. Participants were compensated with course credit or a $10 payment. Six students (9%) were psychology majors; a further 38 students (59%) were majoring in another STEM discipline. Forty-one students (64%) had previously taken a college-level statistics course and 14 students (22%) had previously taken a college-level research methods course.

Materials

Forty-one vignettes and open-ended prompts were created for Study 1 using the following procedure. The vignettes described hypothetical scenarios faced by a researcher, science consumer, or other stakeholder. For example, a scenario could describe a researcher engaging in a specific questionable research practice in order to in-

1 Our preregistration, materials, analysis code, and data can be found on the OSF project repository at https://osf.io/vq26u/.

2 We planned to recruit up to 100 participants for Study 1, but given the novel materials and qualitative nature of the study it was difficult to anticipate how many responses would be necessary to provide a representative sample of students' responses. We conducted a preliminary coding of responses with a sample size of 43 participants and found that responses fell into consistent categories, permitting the creation of target and distractor options. We then decided that a sample of approximately 60 would be adequate for generating the multiple-choice options of the OSCI.

crease their likelihood of obtaining a significant result during statistical analysis.

Each vignette was followed by a question prompting respondents to express their opinion about what that person should do in the situation or if they agreed with the researcher's actions. Prompts were designed to have a similar logical and syntactic structure, following the format: "*Do you agree with how X [...]? Why or why not?*", or "*Would you advise X to [...]?. Why or why not?*".

The full set of items used to prompt open-ended responses in Study 1 can be found at our OSF repository for the project (https://osf.io/vq26u/).

Procedure for Vignette Development

First, we identified a list of target concepts related to open science and questionable research practices (e.g., publication bias, p-hacking, HARKing). We did so by identifying highly-cited recent papers (e.g., John et al., 2012; Simmons et al., 2011; Simonsohn et al., 2014) related to the replicability crisis, and identified questionable research practices and solutions advocated by the open science movement. Although at the time we did not refer to community-generated efforts in this domain, such as the glossary generated by the Framework for Open and Reproducible Research Training (FORRT, Parsons et al., 2022), we did generate a representative subset of those concepts.[3]

Next, we created 41 vignettes addressing these concepts. Of these, 9 items addressed primarily p-hacking (optional stopping, outlier exclusion, dropping participants or observations, covariate inclusion, outcome switching, selective reporting of conditions, outcomes, and experiments), 6 items addressed the incentive structures of science (including publication bias), 5 items addressed replication or computational reproducibility, 5 items addressed misconduct, fraud, or data fabrication, 4 items addressed preregistration, 4 items addressed the use of open data, 2 items addressed HARKing, and 6 items addressed other concepts related to transparency (interrater reliability, participant bias, salami slicing, data peaking). Appendix A shows the distribution of these concepts across items.

Because we aimed to evaluate the understanding of these concepts, we created vignettes that described scenarios faced by individuals from different perspectives (see an example in Table 1). Perspectives included researchers at different career stages and situations (e.g., honors thesis student, graduate student, early career researcher, established researcher, co-author), as well as science consumers and evaluators (e.g., employee of pharmaceutical company, conference paper reviewer, editor), and other stakeholders (e.g., funding agency). Because our goal was to develop an assessment of how people reason about the *application* of concepts related to open science in diverse contexts (not just their factual, definitional knowledge of these concepts), it was necessary to describe the scenario with some level of detail across a few sentences in order to support a preferred path of action.

To improve the readability of our vignettes, during this early design phase, we sought feedback on an initial set of vignettes ($N = 14$) from two experts in academic assessment from the Office of Assessment and Accreditation (OAA) at our institution. Through our discussions, we identified ways to reduce the cognitive load associated with reading the vignettes and to minimize ambiguity in their interpretation. Changes in our vignettes and survey involved the following: (a) presenting a set of definitions of key terms (e.g., hypothesis, sample size, statistical significance, preregistration, peer review) at the beginning of the survey, in order to provide necessary background while offloading the verbiage in individual vignettes, and (b) ensuring that the prompts following the vignettes had a similar syntactic and logical structure (namely, "*Do you agree with how X [...] ? / Would you advise X to [...] ? Why or why not?*)". Toward this end, in some cases, we split a scenario into two (or more) decision dilemmas faced by the researcher, with separate prompts corresponding to each decision point. We also obtained feedback from the experts from the OAA at a second time point, on the entire set of 41 vignettes. This second round of feedback resulted in small edits to improve clarity.

Design and Procedure

The study took place in a laboratory and the survey was administered through Qualtrics. Upon giving consent, participants provided demographic information (age and gender) and information about their academic background, including their major field of study, and whether they had completed a college-level statistics and research methods courses.

At the beginning of the survey, participants were presented with a set of definitions of key terms which appeared in multiple vignettes (e.g., hypothesis, sample size, statistical significance). As noted, this was done to ensure that participants, regardless of their course of study, had the necessary background on these key concepts while minimizing the use of definitions within the vignettes' text. Participants had unlimited time to study these terms.

Next, participants responded to the open-ended prompts of the vignettes. Because initial piloting indicated that the full survey with the 41 items took up to two hours

3 Our 41 initial vignettes represented 23 terms of the FORRT glossary. Note that the FORRT glossary encompasses a number of terms that are beyond the scope of our inventory that include advanced statistical concepts (e.g., Bayes Factor, multiverse analysis), specific open science efforts or scientific tools (e.g., ReproducibiliTea, G*Power), or theoretical perspectives that inform open, transparent, and equitable science (e.g., positionality, intersectionality). In turn, our set of concepts included more specific classifications of FORRT glossary terms that had a single entry (e.g., p-hacking; we considered optional stopping, outlier exclusion, dropping participants or observations, covariate inclusion, outcome switching, selective reporting of conditions, outcomes, and experiments). We also covered concepts related to misconduct, fraud, and data fabrication as they are related to the incentive structures of science and the replicability crisis. These and other terms we represented in the OSCI (e.g., participant bias) are not part of the FORRT glossary.

Table 1. Example vignette from Study 1 and 2.

Item 1 scenario: David's research project is based on a well-known effect in the psychology literature. After attempting to replicate the effect in two experiments, David finds that he hasn't replicated the published findings despite using the same procedure and a large sample of participants. Concerned that he won't be able to publish non-significant results in a journal, he's considering abandoning the project.
Question prompt in Study 1 *Would you advise him to abandon the project or not? Why or why not?*
Sample Responses from Study 1 under emerging themes
Theme: David should still try to publish the results of his project because non-significant findings are informative.
P64: I would not advise David to abandon this project, as it could actually be evidence that the well-known effect in psychology literature might not be so accurate after all. There have been several instances of famous historical psychology experiments dominating discourse for decades, only to have been identified as fraudulent, years later, when researchers like David repeatedly found that the results could not be replicated.
P2: I would advise David to ask a professor or peer about his findings and ask for their advice. Non significant results are actually significant results because it shows error with the initial study that was being replicated. […] If someone else also finds that they cannot reproduce the study, then the results from David should be published.
Theme: David should keep modifying the procedure until he obtains a significant effect that he can then publish.
P4: Abandon it and go about it in a different way and keep in mind the first two experiments. I think it's important to move forward and not stay stuck on something deemed nonsignificant
Theme: David should not try to publish these results because replicating someone else's work is unethical
P48: I would advise him to abandon the project as it is already a published project, replicating someone else's work and trying to publish it in a journal is nonethical and cheating.
Theme: David should not try to publish these results because non-significant results are not informative.
P4: Abandon it and go about it in a different way and keep in mind the first 2 experiments. I think it's important to move forward and not stay stuck on something deemed nonsignificant.
Question Prompt in Study 2 *What should David do?*
Multiple Choice Options for Study 2
a. David should still try to publish the results of his project because non-significant findings are informative. (best response)
b. David should keep modifying the procedure until he obtains a significant effect that he can then publish.
c. David should not try to publish these results because replicating someone else's work is unethical.
d. David should not try to publish these results because non-significant results are not informative.

Note: This example illustrates how open-ended responses from participants in Study 1 were adapted into multiple choice options in Study 2.

to complete, items were presented in one of two lists (of 20 and 21 items respectively) to shorten the duration of the study and reduce participants' fatigue. Items across the two lists were matched to the extent possible in terms of the topics covered, prompt complexity, and number of multi-part questions represented (where different questions are associated with the same scenario). Appendix A lists the distribution of topics across the items of the two lists. To mitigate fatigue and other order effects, within each list, the presentation of the items was randomized with the constraint that multi-part questions were presented in sequence. An equal number of participants completed each list of items (32 participants per list).

After responding to the vignettes, participants were debriefed about the purpose of the study and were compensated for their participation. Sessions took approximately one hour to complete.

Results and Discussion

The results of Study 1 were qualitative. The final product was a set of multiple-choice options for each question item, based on participants' open-ended responses. To generate these multiple choice questions, we used the following procedure: (1) Before viewing participants' responses for a given vignette, we generated a target response for each item corresponding to the best course of action. This was done to establish whether at least a subset of the participants could produce the target response spontaneously. (2) We grouped participants' open-ended responses under consistent themes. When two or more responses made the same point, we grouped them under that common theme. If a participant's response addressed more than one theme, we cross-listed the response under all relevant groups, with the pertinent phrases of the response highlighted in each group. If a single responder made a distinctive well-reasoned point that could serve as a distractor option, we listed it as a separate theme. We grouped generic responses together (e.g., "because it's biased" or "this may skew the results"), along with other miscellaneous or uncodable responses that did not fit into any of the emerging themes. (3) We summarized each theme in a statement capturing the common thread of the responses in that group (e.g., "I would advise David to not abandon the project because non-significant findings are informative and should be published" in response to the scenario in Table 1). Whenever possible, we used language similar to that used by participants. (4) Finally, we referred to the target response we

generated in step 1 to identify the correct response statement among the summary statements for the groups of responses. If needed, we revised the wording of the target response to capture more accurately the conceptual distinction that we wanted students to make when reasoning about the vignette

Once we had generated summary statements for the themes of responses in each vignette, we selected and refined a set of four multiple-choice options to be used in Study 2 as follows: (1) We included the correct / best response. (2) We included response themes that were clearly incorrect (e.g., "David should not try to publish these results because replicating someone else's work is unethical"). If necessary, we modified the wording of these statements, which was based on participants' open-ended responses, to minimize ambiguity in their interpretation. (3) We excluded response themes that were broadly accurate but did not represent the target response for the prompt (e.g., "I would advise David to not abandon the project because additional replications are needed to establish if there is a true effect."). The rationale was to select multiple-choice options that clearly distinguished between correct and incorrect responses. (4) If, as the result of this process, there were fewer than 3 response options that could be included as distractors, we generated additional options to yield a total of 3 distractors. (5) Where necessary, we made revisions to the vignettes. For some vignettes, step 3 revealed that participants produced responses that were broadly correct due to ambiguity in the scenario or in the prompt. In those cases, we slightly modified those vignettes to minimize ambiguity about the preferred course of action in that scenario.

Three coders (the two authors and a research assistant) independently completed the qualitative coding of different subsets of items, grouping participants' open-ended responses into themes and generating a preliminary set of multiple-choice options. Coders then jointly discussed all coded items and selected the final set of multiple-choice options for Study 2. One item was excluded because coders could not identify sufficiently distinctive correct and incorrect options.[4] This yielded a final set of 40 vignettes along with their multiple-choice options to be used for Study 2. Nine items were part of multi-item groups based on a common scenario (items 7–9, 10–11, 12–13, and 14–15).

Study 2

The goal of Study 2 was to evaluate the multiple-choice version of the OSCI in a larger sample of undergraduates. The full set of 40 multiple-choice items were presented to a new group of participants. Responses were then analyzed with Item Response Theory (IRT) to select the final set of items for the OSCI.

Participants

N = 262 undergraduate students (149 identified as female, 113 as male) at UNC Charlotte in the United States participated for course credit or for a $10 payment. A further *N* = 6 individuals were excluded from analysis due to incomplete responses. The mean age was 20.46 years (*SD* = 4.28). We did not collect information about the participants' race and ethnicity. In our preregistration, we had planned to collect data from 200 participants, based on a prior study that developed a concept inventory for research methods (*N* = 208, Veilleux & Chapman, 2017). However, because of participant availability, we ended up running an additional 62 participants.[5]

Thirty-seven students (14%) were psychology majors, while a further 101 students (39%) were majoring in another STEM field. 84 students (32%) had previously taken a college-level statistics course, 12 students (5%) had previously taken a research methods course but no statistics course, and 44 students (17%) had taken both a statistics course and a research methods course. The remaining 122 students (47%) had not taken either statistics or research methods courses at the college level.

Materials and Procedure

We used the set of 40 vignettes and their multiple-choice response generated and selected through the process described in Study 1. This set of items can be found at our OSF repository for the project (https://osf.io/mxv59/).

As with Study 1, the study took place in a laboratory. The survey was administered using Qualtrics and was set up similarly to Study 1. Participants first provided demographic information (age, gender) and information about their major field of study, and their statistics and research methods experience. Next, participants were presented with the same review of "Key Concepts" as in Study 1. Participants had unlimited time to study these concepts and were not tested on them directly before proceeding to the vignettes. Participants then responded to the 40 vignettes associated with multiple-choice response options. Items were presented in random order, except that multipart questions were presented in sequence. At the end of the study, participants were debriefed and compensated for

4 This scenario was about a Ph.D. student whose dilemma was whether to learn programming for statistical analyses, given that this time investment could delay their degree progress. This scenario did not lend itself to a good contrast between "correct" and "incorrect" courses of action, and was thus excluded.

5 Because Study 2 was the first time we assessed performance on the OSCI items and we had no prior knowledge about their properties, we expected that a formal power analysis would have been associated with too much uncertainty to provide guidance on a target sample size. While there are no definitive standards for sample size requirements when evaluating IRT models, recent work suggests that samples of 250 are often sufficient when comparing models (Zimmer et al., 2022) or for the purpose of parameter estimation (Finch & French, 2019) for a similar number of items and the 2PL model as used in our study.

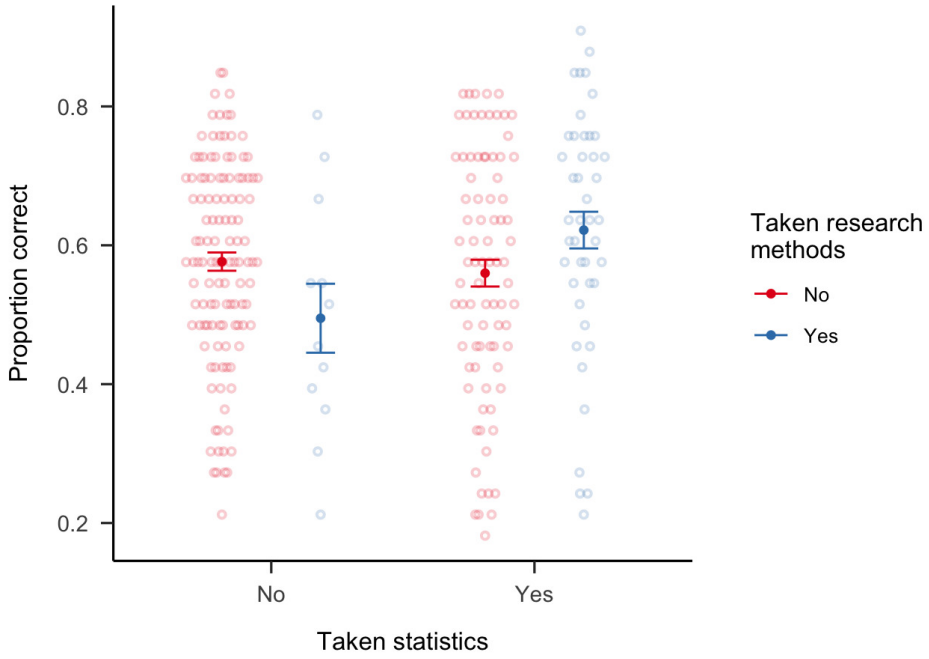

Figure 1. Proportion of correct responses on the refined subset of 33 OSCI items based on prior coursework on statistics and research methods in Study 2.

Error bars indicate standard errors of the mean.

their participation. Sessions took approximately one hour to complete.

Results

Responses were treated as dichotomous (correct/incorrect) according to the best responses to each vignette identified in Study 1. Participants selected the correct response on 22.34 items on average (*SD* = 5.91), an average proportion of 56% correct (*SD* = 15%). The proportion of correct responses (see Table 2) ranged from 16–87%, indicating that the items covered a range of difficulty levels.

IRT analysis. We used Item Response Theory (IRT; Toland, 2014) to model responses to the OSCI questions. IRT models were fit using the LTM R Package (Rizopoulos, 2006). IRT models responses to multi-item assessments as a function of item difficulty and individuals' latent ability in the domain. In our preregistration we planned to estimate a 3-parameter model (3PL) which includes item-specific parameters for discrimination, difficulty, and the likelihood of guessing. Discrimination (*a*) indicates the extent to which an item distinguishes between people of varying ability levels, captured by the slope of the item characteristic curve (ICC). Difficulty (*b*) indicates the ability level associated with a 50% chance of responding correctly. Given the large number of parameters in the 3PL model relative to our sample size, we departed from the preregistration by comparing it to simpler IRT models: the 2-parameter model (2PL) which includes item-specific parameters for discrimination and difficulty; and a 1-parameter (Rasch) model with item-specific parameters for difficulty and a single discrimination parameter for all items. The 2PL model provided a better fit of the data compared to the 1-parameter (Rasch) model (likelihood ratio test, $\chi^2(39) = 246.9$, $p < .001$), while the 3PL model did not significantly improve performance beyond the 2PL model ($\chi^2(40) = 40.4$, $p = .45$).

Table 2 lists the results from fitting the 2-parameter (2PL) model for the initial 40 items evaluated for inclusion in the final OSCI. As per our preregistration, we excluded: 1 item due to a negative point biserial correlation (item 12), since this indicates that students who answered this item correctly did poorly on the overall test; another item was excluded because it appeared in the same question group (item 13); 2 items due to a significant χ^2 test indicating poor fit to responses (items 11 and 32); and 3 items due to low or negative discrimination parameters (items 26, 27, 34). Excluding these 7 items led to a final set of 33 items.

Relationship to prior experience. Following our preregistration, we conducted planned exploratory analyses to examine whether performance on the refined set of 33 items of the OSCI was related to prior experience, as measured by previous enrollment in statistics and research methods courses (Figure 1). In separate models, regression was used to model the raw scores (proportion of correct responses) and the predicted ability levels based on the 2PL model described in the previous section, using only those items that were retained after exclusions. Since similar results were obtained for proportion correct and estimated ability, we only report results for proportion of correct responses for ease of interpretation (for the results on estimated ability see supplementary material on OSF: https://osf.io/yrdng/). A logistic regression model of the proportion of correct responses indicated a negative effect of having taken research methods and a significant interaction between having taken both statistics and research

Table 2. Results of IRT analysis in Study 2. Items that were excluded from the final OSCI are highlighted.

			IRT parameters (coefficient and 95% CI)		
Item	% correct	Point Biserial	Difficulty (*b*)	Discrimination (*a*)	*p*-value for χ^2 test
1	0.64	0.53	-0.52 [-0.76, -0.29]	1.73 [1.17, 2.28]	0.114
2	0.55	0.07	-1.95 [-8.03, 4.14]	0.09 [-0.18, 0.37]	0.086
3	0.53	0.31	-0.16 [-0.52, 0.2]	0.78 [0.45, 1.11]	0.333
4	0.79	0.3	-1.67 [-2.32, -1.01]	0.93 [0.52, 1.35]	0.641
5	0.52	0.26	-0.1 [-0.49, 0.29]	0.69 [0.37, 1.01]	0.653
6	0.79	0.43	-1.21 [-1.57, -0.86]	1.51 [0.97, 2.05]	0.461
7	0.59	0.36	-0.41 [-0.68, -0.13]	1.2 [0.78, 1.62]	0.238
8	0.23	0.04	6.86 [-5.4, 19.12]	0.18 [-0.14, 0.51]	0.525
9	0.48	0.33	0.09 [-0.21, 0.39]	0.99 [0.62, 1.36]	0.18
10	0.87	0.47	-1.41 [-1.71, -1.11]	2.58 [1.51, 3.65]	0.535
11	**0.78**	**0.41**	**-1.16 [-1.5, -0.83]**	**1.59 [1.02, 2.15]**	**0.018**
12	**0.16**	**-0.02**	**-23.57 [-146.87, 99.73]**	**-0.07 [-0.43, 0.29]**	**0.134**
13	**0.52**	**0.18**	**-0.28 [-1.02, 0.47]**	**0.35 [0.06, 0.63]**	**0.663**
14	0.63	0.23	-1 [-1.66, -0.34]	0.58 [0.26, 0.89]	0.168
15	0.83	0.33	-1.87 [-2.56, -1.17]	1.03 [0.57, 1.49]	0.375
16	0.69	0.42	-0.83 [-1.14, -0.51]	1.26 [0.81, 1.7]	0.91
17	0.75	0.33	-1.22 [-1.65, -0.79]	1.14 [0.7, 1.58]	0.679
18	0.35	0.09	4.81 [-6.21, 15.83]	0.12 [-0.16, 0.41]	0.393
19	0.61	0.16	-1.32 [-2.62, -0.02]	0.34 [0.05, 0.63]	0.539
20	0.69	0.36	-1.02 [-1.47, -0.58]	0.94 [0.56, 1.32]	0.641
21	0.79	0.41	-1.18 [-1.51, -0.84]	1.6 [1.03, 2.18]	0.202
22	0.71	0.31	-1.24 [-1.78, -0.7]	0.86 [0.49, 1.23]	0.928
23	0.29	0.21	2.1 [0.55, 3.66]	0.43 [0.12, 0.75]	0.671
24	0.5	0.33	0.01 [-0.31, 0.34]	0.88 [0.53, 1.22]	0.349
25	0.23	0.22	2.04 [0.92, 3.17]	0.63 [0.27, 1]	0.888
26	**0.35**	**0.02**	**-0.75 [-8.39, 6.89]**	**0.04 [-0.23, 0.31]**	**0.19**
27	**0.51**	**0.06**	**-373.85 [-6.62e4, 6.53e4]**	**0 [-0.28, 0.28]**	**0.529**
28	0.48	0.25	0.13 [-0.31, 0.58]	0.6 [0.29, 0.91]	0.387
29	0.85	0.25	-2.45 [-3.6, -1.29]	0.82 [0.37, 1.26]	0.196
31	0.57	0.26	-0.44 [-0.82, -0.05]	0.77 [0.44, 1.11]	0.333
32	**0.73**	**0.22**	**-1.61 [-2.42, -0.8]**	**0.68 [0.33, 1.03]**	**0.038**
33	0.48	0.3	0.09 [-0.28, 0.45]	0.75 [0.42, 1.08]	0.683
34	**0.30**	**0.01**	**-11.15 [-54.62, 32.33]**	**-0.08 [-0.37, 0.22]**	**0.858**
35	0.46	0.2	0.39 [-0.23, 1.02]	0.44 [0.15, 0.73]	0.948
36	0.55	0.16	-0.99 [-2.64, 0.66]	0.22 [-0.06, 0.5]	0.088
37	0.47	0.25	0.23 [-0.2, 0.66]	0.63 [0.32, 0.94]	0.088
38	0.69	0.39	-0.89 [-1.25, -0.54]	1.14 [0.72, 1.56]	0.982
39	0.24	0.14	3.72 [-0.02, 7.46]	0.32 [-0.01, 0.65]	0.327
40	0.65	0.26	-0.88 [-1.35, -0.4]	0.78 [0.43, 1.12]	0.844
41	0.48	0.22	0.13 [-0.31, 0.57]	0.6 [0.3, 0.91]	0.998

Note. Items 11 and 32 had significant χ^2 test indicating poor fit to responses. Item 12 had a negative point biserial correlation. Items 26, 27, and 34 had low or negative discrimination parameters (a). Item 13 was excluded because it involved the same scenario as item 12.

methods (Table 3). We used linear contrasts to compare the predicted accuracy for each combination of the statistics and research methods factors (using the *emmeans R library,* Lenth et al., 2019, with the Tukey correction for multiple comparisons). Compared to students who had taken both statistics and research methods, scores were lower among students who had taken neither (Odds ratio (OR) = .83, 95% CI [.71, .98], $z = -3.01$, $p = .01$), students who had only taken research methods (OR = .60 [.44, .80], $z = -4.54$, $p < .001$), and students who had only taken statistics (OR = .77 [.65,

Table 3. Results of logistic regression model for proportion correct in Study 2.

	Estimate	SE	2.5%	97.5%	*z*	*p*
(Intercept)	0.308	0.032	0.246	0.371	9.67	< .001
Stats	-0.068	0.05	-0.165	0.03	-1.36	0.174
Research Methods	-0.329	0.105	-0.536	-0.122	-3.117	0.002
Stats x Research Methods	0.586	0.125	0.341	0.83	4.702	< .001

.92], $z = -3.88$, $p < .001$). Among those students who had not taken research methods, there was no difference based on whether they had taken a statistics course or not (OR = 1.07 [.94, 1.22], $z = 1.36$, $p = .52$). Surprisingly, among those who had not taken statistics, scores were higher for those who had not taken research methods compared to those who had (OR = 1.39 [1.06, 1.82], $z = 3.12$, $p = .01$). Although this implies a negative effect of research methods coursework on performance, it should be noted that this group (who had taken research methods but not statistics) included only a dozen students and this finding should be interpreted with caution.[6]

Discussion

The goal of Study 2 was to test the multiple-choice OSCI in a large group of undergraduates and evaluate its psychometric properties using an item response theory analysis. Upon evaluating the IRT parameters of the original set of 40 items, we refined the OSCI to a final set of 33 items. In addition, we obtained initial evidence that performance on the concept inventory was related to students' prior training, as scores on the final set were highest among students who had previously taken both college-level statistics and research methods courses. These differences based on prior coursework could be due to a number of factors. For example, in these courses, students gain greater familiarity with research- and statistics-related concepts that could help them understand and reason about the scenarios. Another possibility is that students experienced direct exposure to open science practices in the context of these courses. Self-selection bias could have also played a role: students who had taken these courses could have more aptitude to reason about science. Having obtained this initial evidence for the OSCI's sensitivity to prior coursework, we set out to evaluate whether it could capture learning gains associated with the introduction of a targeted curriculum in a psychology research methods course.

Study 3

Study 3 was a preliminary "proof of concept" for using the OSCI as an assessment tool. The final set of 33 items selected for the OSCI were used to evaluate student learning across two implementation rounds in a research methods course in Fall 2019 and Spring 2020. Learning gains were evaluated using a longitudinal design with two timepoints: a pre-test prior to the introduction of the target curriculum in the course and a post-test at the end of the semester. Along with the OSCI, an additional questionnaire was administered to assess changes in students' self-efficacy and broader attitudes toward psychological research.

Participants

Forty-seven undergraduate students majoring in Psychology enrolled in a research methods course at UNC Charlotte in the United States participated for extra credit for that course. Participants came from four different sections of this course taught across two semesters (Fall 2019 and Spring 2020). The first and second author each taught one of these sections per semester. The potential sample size was limited by the enrollment cap of the course (25 students per section). Of the 47 participants who enrolled in the study, 37 completed both the pre-test and post-test (19 in Fall 2019, 18 in Spring 2020).[7] In the final sample

6 Based on a reviewer's suggestion we explored whether age was a confounding factor in these results, since more advanced students would naturally be expected to have experienced a broader range of courses. We found that the average age was indeed higher among participants who had taken both stats and research methods (M_{age} = 22.1 years) compared to participants in the other groups (M_{age} = 20.1, 19.7, and 20.3 years). Repeating the main analysis of performance with age included as a covariate indicated a small but significant effect of age on proportion correct (β = .01 [.003, .02], $z = 2.59$, p = .01). In this model there was no effect of stats (β = -.06 [-.16, .04], z = -1.25, p = .21), but there was a significant negative effect of having taken research methods (β = -.32 [-.53, -.12], z = -3.06, p = .002) and a significant interaction (β = .55 [.30, .79], $z = 4.37$, p < .001), all effects which are consistent with those of the model reported in the main text.

7 We compared performance on the pre-test between groups of participants who did or did not go on to complete the post-test. A logistic regression analysis of proportion correct indicated higher accuracy among those participants who did not complete the post-test (M = .65, SD = .21) compared to participants who completed both sessions (M = .54, SD = .16), Wald z=13.19, p< .001. This could be because those who dropped out of the study were high-performing students in good standing in the course, who by the end of the semester did not need the additional extra credit associated with the post-test. We do not explore this possibility further, as examining the relationship between the students' academic performance (through course grades or GPA) and their performance of the OSCI is beyond the scope of analyses covered by the IRB protocol for this work.

there were 29 students identifying as female and 8 as male; their mean age was 21.59 years (*SD* = 4.15). We did not collect information about the participants' race and ethnicity. Given the relatively small sample size we obtained, we deviated from our original preregistration (where we had proposed analyzing data from two semesters separately), and pooled the data of students from two semesters into one sample for analysis.

Materials

Open Science Concept Inventory (OSCI)

The entire set of 40 items of Study 2 was administered to students; however, the focus of Study 3 is on the retained 33 items of the OSCI selected in Study 2.[8]

Open Science Curriculum

The new curriculum on open and transparent research practices included three recorded video lectures.[9] The student learning objectives of this curriculum were aligned with the targeted development of the OSCI. The objectives were that students could recognize the problems with the pervasive incentive structures in science, distinguish between research misconduct and QRPs, recognize the effects of common QRPs on robustness of scientific literature, identify open science solutions to address these problems, think critically about research practices, and be able to situate open science guidelines within the scientific ecosystem (including funding agencies and academic publishers). The video lectures introduced the credibility crisis in Psychology (part 1a), questionable research practices (part 1b) and open science solutions (part 2). The topics covered in the three videos are described in more detail in Appendix B. The video lectures and lecture slides can be accessed at our OSF repository.

Self-efficacy and Attitudes Toward Research

We developed a questionnaire on self-efficacy and attitudes toward psychology research by taking questions from Kardash (2000) and Chopik et al. (2018) (see Appendix D for the full set of questions). Our analyses focused on items that were related to three broad constructs: self-efficacy (10 items), interest in research (4 items), and trust in psychological research (5 items). Self-efficacy was measured by having students rate their perceived ability to complete certain research activities (e.g., generating a hypothesis, statistical analysis, etc.). Interest in research involved questions about students' own motivation to pursue research-related careers. Trust in psychological research was measured through students' degree of agreement with statements about the results of psychological studies. Responses to all items were made on a 5-point scale (from 1 = "Strongly disagree" to 5 = "Strongly agree"). We also included items from Chopik et al. (2018) which assessed overall attitudes toward psychology (items 1–5), but did not analyze them further since there was no clear connection between the open science curriculum and broad attitudes about psychology or its relationship to other fields.

Procedure

To minimize the likelihood that students enrolled in the course felt coerced to participate, each instructor made the announcement about this study at the beginning of the semester as an extra credit opportunity in the other instructor's class. Students were informed that their participation was optional and that their instructor would not be informed about whether they decided to participate until after the course was complete. Participants received 1% extra credit toward their course grade for completing the pre-test and an additional 2% for completing the post-test.

The pre-test was administered after the instructors completed an overview of statistics in the course, which included core concepts such as null hypothesis testing, Type I and Type II error, and statistical power. This ensured that participants had the necessary background on these key constructs. Students who volunteered to take part in the study were sent a link to a Qualtrics survey and had one week to complete it.

After launching the survey, participants provided informed consent and responded to demographic questions (age and gender). The format of the survey was the same as Study 2: participants were presented with a list of "Key Concepts", followed by the 40 final items of the OSCI with multiple choice options. After responding to the OSCI items, participants completed the questionnaire for self-efficacy, interest, and attitudes toward research.

The open science module was available on the learning management system for the courses between weeks 7 – 15 of the semester. Students typically had a week to view each of the three videos and answer the questions of a low-stakes quiz embedded within the video. These quiz questions were inserted to ensure that the students paid attention; scores counted towards the students' assignments grade for the course and was independent from extra credit for participating in Study 3.

After the open science curriculum had been completed, students were sent a link to the Qualtrics survey for the

8 We administered all 40 items because Studies 2 and 3 were run concurrently in Fall 2019 and early Spring 2020. During data collection, we did not know which items would be retained to assess learning gains in the classroom in Study 3, until conducting the IRT analyses for Study 2.

9 The curriculum was also supplemented with hands-on labs, which were modeled on activities from OSL Stats Labs (https://sites.google.com/view/openstatslab/home) and focused on reproducing published results using open datasets and the open-source statistical platform R. These labs are available on our OSF repository.

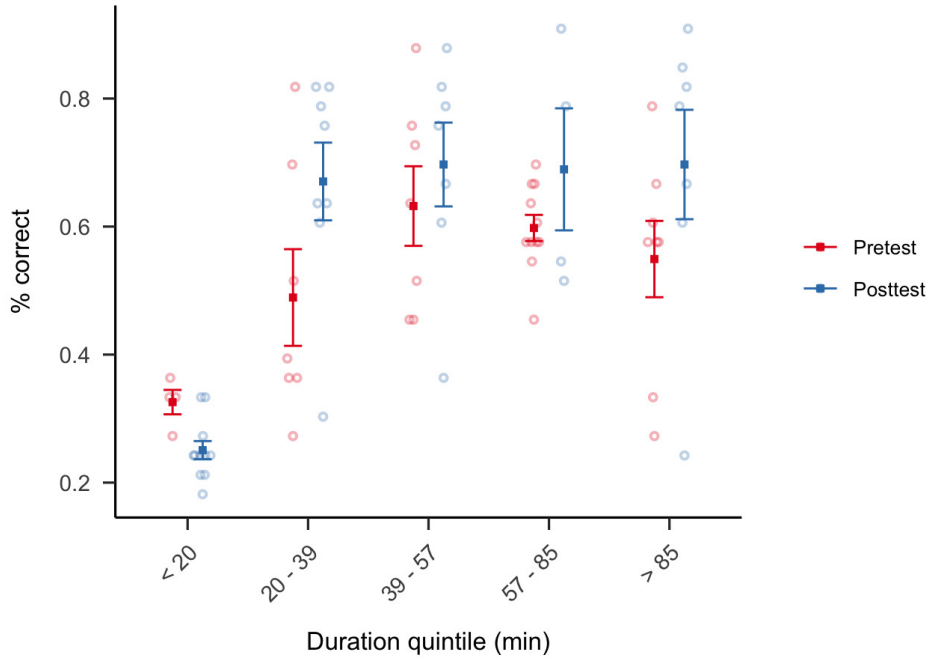

Figure 2. Proportion correct in Study 3 on the Pre- and Post-test after splitting into quintiles based on task duration.

Error bars indicate standard errors of the mean.

post-test. Participants had one week to complete the post-test, which was identical to the pre-test.

Results

Accuracy. Proportion correct was comparable at the Pre-test (Fall 2019: M = .60, SD = .12; Spring 2020: M = .49, SD = .19) and the Post-test (Fall 2019: M = .61, SD = .24; Spring 2020: M = .50, SD = .26). For proportion of correct responses, there was no effect of timepoint (β = .03, SE = .04, 95% CI [–.05, .12], z = .75, p = .46), semester (β = .24, SE = .12, 95% CI [–.04, .51], z = 1.73, p = .08) or instructor (β = .05, SE = .14, 95% CI [–.23, .33], z = .37, p = .71). Similar results were obtained for predicted accuracy scores derived from the IRT model (see Supplementary Materials: https://osf.io/yrdng).

During the analysis we observed some surprisingly low task durations given the number of items to be completed, suggesting low-effort responses from some participants (minimum duration: 3.5 minutes). We therefore conducted a non-preregistered, exploratory analysis to examine whether the amount of time taken to complete the task was related to performance. Total task durations for both the Pre- and Post-test were aggregated and binned to form equal-sized quintiles. Figure 2 shows proportion correct across task duration quintiles for the Pre- and Post-test. There was a marked difference in performance among the lowest quintile compared to the other groups, with participants who completed the task in less than 20 minutes performing near chance levels at both the Pre- and Post-test (and performing better at Pre-test, in fact). Performance was higher among participants with longer task durations and appeared to increase from Pre-test to Post-test. This was confirmed in a separate logistic regression model after excluding any participants who fell in the lowest quintile for either the Pre- or Post-test (N = 12, 4 in Fall 2019, 8 in Spring 2020). This indicated an effect of timepoint (Table 4), with proportion correct lower in the Pre-test (Fall 2019: M = .62, SD = .08; Spring 2020: M = .60, SD = .17) than the Post-test (Fall 2019: M = .71, SD = .16; Spring 2020: M = .66, SD = .22). There were again no effects of semester or instructor for this subset of participants. The findings were similar for the students' estimated ability, for which we also documented a significant effect of time point (see Appendix E).

Qualitative examination of learning gains

Given that students who expended more effort when completing the OSCI did improve in their performance, we qualitatively examined item-based performance to identify areas of the curriculum that were particularly effective. The items of the OSCI that showed the greatest improvement in accuracy (⩾ 20%) concerned the following topics: the value and informativeness of null findings (e.g., when reviewing papers, item 2; when deciding to continue data collection after piloting, item 33), the value of exact replications and not just novel extensions (item 37), the weighing of evidence from preregistered versus non-preregistered studies (item 3), the appropriateness of adhering to one's original analysis plan (e.g., to exclude outliers, item 25), the importance of transparently reporting exploratory and confirmatory analyses (item 24), and recognizing instances of HARKing (item 8). Improvement on these items, collectively, could indicate that students recognized the value of replication, the distinction between confirmatory and ex-

Table 4. Results of logistic regression model for proportion correct in Study 3.

	Estimate	SE	2.5%	97.5%	*z*	*p*
(Intercept)	0.661	0.133	0.392	0.937	4.954	< .001
Timepoint	0.179	0.054	0.074	0.285	3.326	0.001
Semester	0.065	0.133	-0.208	0.336	0.491	0.624
Instructor	0.057	0.131	-0.211	0.326	0.437	0.662

ploratory analyses, and the importance of being transparent about any departures from analysis plans. Such improvements make sense given that these items map onto the central themes of the three video lectures: with the first lecture highlighting that the robustness of findings can be established through repeated replications, the second lecture highlighting problematic departures from initial plans (including p-hacking and HARKing), and the third lecture delineating pathways to more transparent research.

The shift from Pre-test to Post-test in the distribution of correct versus competitor options that students chose in many of the OSCI items also demonstrates these learning gains. For example, for the item about HARKing, at pre-test the most prevalent response (16/37 responses; 43%) was that it was acceptable for the researcher to reframe his original hypothesis (that liberal attitudes predict depression) after conducting an exploratory analysis with a subset of respondents (from non-coastal states), because "he reported clearly that his findings only concern non-coastal states". At Post-test, only 19% of students (7/37) chose that response. The most prevalent response at post-test was the correct one, indicating recognition of HARKing ("I do not agree with how Dr. Smith presented the results because excluding coastal states was not part of his original analysis plan or hypothesis." 14/37 students; 38%).

We also considered performance on those items that addressed practices associated with p-hacking, such as outcome switching (item 17), optional stopping (item 4), and selectively reporting experiments (item 21), conditions (item 20), and outcome measures (item 19). Students exhibited relatively high accuracy at Pre-test (ranging from 72-80%) and numerically similar performance at Post-test (ranging from 72-84%). Since students in this course had taken statistics and another research methods class as prerequisites, this prior background may have allowed them to reason their way through these scenarios. To achieve greater improvement on this set of concepts around p-hacking, it might be necessary to integrate the curriculum more closely with other elements of the course, including in-class discussions and labs on data analysis.

Trust, interest, and self-efficacy. Finally, we examined any changes in trust, interest, and self-efficacy from the pre-test to post-test. Average responses to all items on the attitudes questionnaire are shown in Appendix D. In accordance with our preregistration, we calculated mean scores for items related to trust in the robustness of psychology (items 6–10), interest in pursuing careers in research (items 11–14), and self-efficacy in engaging in research (items 15–24). Each outcome was modeled with mixed effects linear regression with timepoint (Pre/Post), semester (Fall 2019, Spring 2020), and instructor as fixed effects, with random intercepts for participants. Because the preceding analysis suggested low-effort participation among some students, here we report the results after excluding participants from the lowest task duration quintile. However, the results are not meaningfully different when these analyses are performed without any exclusions, in line with our preregistration (see Supplementary Material for results with the full sample).

There was a significant effect of timepoint on self-efficacy (β = .19, SE = .06, 95% CI [.07, .31], $t(24)$ = 3.06, p = .005), with self-efficacy higher at the Post-test (M = 4.18, SD = .45) than the Pre-test (M = 3.80, SD = .57). Cronbach's alpha for self-efficacy was α = .86. Self-efficacy did not differ across semesters (β = .02, SE = .09, 95% CI [–.15, .19], $t(22)$ = .19, p = .85) or instructors (β = –.06, SE = .09, 95% CI [–.23, .11], $t(22)$ = –.70, p = .49). We did not find any significant differences in trust (M = 3.87, SD = .48; Cronbach's α = .64) or interest (M = 4.04, SD = .79; Cronbach's α = .73) across timepoints, semesters, or instructors (all p > .31; see Appendix E).

Discussion

We predicted that the introduction of open science topics would improve students' conceptual understanding of robust and reproducible research methods, as measured by the OSCI. While there were no overall differences in the full sample, we found evidence for improvement from Pre- to Post-test in an exploratory analysis that excluded students who appeared to exhibit low-effort participation based on their short task durations. The initial non-significant finding could be an artifact of low effort responders, who in fact performed better at the Pre-test than the Post-test. This apparent low-effort participation, observed in a considerable portion of the students (32% of the sample), should be contextualized in part by the COVID-19 pandemic: the majority of the excluded participants (8 out of 12) completed the study during Spring 2020, a period of major upheavals due to COVID-19, including a mid-semester move to online instruction. Although the transition to online instruction did not directly interfere with the study procedures since the lectures and surveys were delivered online, it may have contributed to disengagement among students.

Learning gains on individual items indicated broad improvements in performance on items that were related to the key themes of the lectures, including the importance of replication and transparency, as well as potential sources of bias that emerge at different points in the research process. These results add to previous findings that brief instruc-

tional interventions can improve understanding of the epistemic foundations of scientific evidence and systemic biases that can undermine the robustness of published research (e.g., Blincoe & Buchert, 2020; Chopik et al., 2018; Sarafoglou et al., 2020). In contrast, there was less evidence of learning gains for items related to p-hacking, particularly when the target concept involved analytic flexibility without obvious signs that the researcher was withholding information (e.g., optional stopping, covariate inclusion, data exclusion). This suggests the need for more in-depth discussion or hands-on activities to help students to recognize the effects of p-hacking and sources of flexibility in analysis procedures.

In addition to our predictions about improved performance on the OSCI, we had also predicted that students would have improved attitudes toward psychology as a field and increased self-efficacy about conducting research, as measured by the questionnaire on attitudes toward research. We found an improvement in self-efficacy but no changes in trust or interest in research. As shown in Appendix D, however, students' responses on several of the items measuring trust (e.g., items 6-8) and interest (e.g., items 11-12) were already high at the Pre-test, suggesting that there wasn't much room for growth on these measures among this sample of psychology majors. Nevertheless, this result echoes previous findings that knowledge gains in psychology research methods are not always accompanied by improved attitudes toward conducting research or perceptions of the utility of research skills (Ciarocco et al., 2013; Holmes & Beins, 2009; Sizemore & Lewandowski, 2009). Overall, students concluded the course with improved understanding of open and transparent research practices and higher self-efficacy about engaging in research.

General Discussion

We set out to develop and evaluate the Open Science Concept Inventory (OSCI), an instrument designed to assess conceptual understanding of open and transparent research practices. We first created a set of scenarios depicting research dilemmas and, in Study 1, elicited open-ended responses about how researchers should respond to each situation. These responses were distilled into multiple choice options that included students' common misconceptions and a "best practice" tied to each concept. In Study 2, we evaluated our initial set of items in a sample of undergraduate students, using an IRT analysis to select 33 items that discriminated between respondents of different ability levels. In that study, the OSCI was sensitive to students' past experience in relevant coursework, such that those who had taken both research methods and statistics outperformed the rest. This finding provided some initial evidence that reasoning about the open science scenarios on the OSCI relies on research- and statistics-related knowledge, though it is unclear to what extent those prior courses touched on issues specific to open science and the replication crisis. Study 3 was a preliminary proof of concept in which we used the OSCI to assess the effectiveness of a targeted curriculum on open science in an undergraduate research methods course, using a pre- / post-test design. Even though we did not observe the predicted changes in performance on the OSCI in our planned analysis with the full sample, there was evidence of learning gains after excluding participants who seemed to complete the study with low effort. Among the remaining students, the largest improvements were related to the central themes of the video lectures, suggesting that students were able to apply the concepts introduced in the materials to the scenarios on the OSCI. In addition to improved performance on the OSCI, these students also reported increased self-efficacy in research, while their interest and trust in psychological research did not change. Although the results of Study 3 should be treated with some caution given their exploratory nature, they provide tentative evidence that the OSCI can measure learning gains arising from a targeted curriculum.

Our findings contribute to pedagogical research in psychology by showing that instruction on open science practices can bolster students' conceptual foundations for research methods, including the importance of replicability in evaluating evidence, the risks of questionable research practices, and the broader systemic biases that distort the scientific literature. We see this work as complementing recent efforts to align research methods instruction with the field's best practices, such as reorienting student projects toward replication (Frank & Saxe, 2012; Wagge et al., 2019). These approaches have clear benefits in providing hands-on, practical training in open and transparent methods, as well as providing students opportunities to contribute to the scientific literature through replication and extension. Although these educational efforts are increasingly common, attempts to systematically assess students' conceptual understanding of open and transparent research practices have been lacking (Pownall et al., 2022). Our results indicate that the OSCI can be a useful pedagogical tool for measuring students' ability to reason about the nature of scientific evidence. Understanding how decisions by different parties can introduce biases at each step along the scientific workflow, and how these biases can be mitigated through open and transparent research practices, is a crucial skill for both future scientists and science consumers.

Integrating the lessons of the past decade into undergraduate education does not come without significant costs in terms of time and effort. We aim to facilitate this by making our materials and assessments openly available on OSF so that they may be easily adapted in other research methods courses. The video lectures and OSCI can stand alone or be integrated into existing curricula for research methods in psychology to support student learning objectives centered on critically evaluating scientific findings, identifying robust research methods, and conducting research using the field's best practices.

Limitations and Future Directions

We sought to include in the OSCI a broad range of concepts that we view as being central to the replication crisis and open science. Unlike more established components of the undergraduate psychology curriculum like frequentist statistics, however, shifts in research practices are ongoing

and likely will continue to evolve, as seen in ongoing debates over the distinction between confirmatory and exploratory research (e.g., Devezer et al., 2019; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019) and the importance of preregistration (e.g., Szollosi et al., 2020). We recognize that any instrument assessing research practices should also evolve and be guided by consensus among researchers about best practices. Future development of the OSCI would benefit from a more systematic, community-driven approach, for example by leveraging a collaborative knowledge base like the FORRT glossary (Parsons et al., 2022) to identify concepts that warrant representation.

Moreover, while we aimed to create a collection of scenarios that were not specific to any one discipline, some of the concepts we selected may be more or less relevant to researchers in specific sub-fields or from different methodological traditions. Most of the focus of research on QRPs has been on quantitative methodologies. Some concepts may also be applicable to qualitative analyses (e.g., lack of transparency in reporting changes to an analysis approach), while other concerns (e.g., p-hacking, statistical power) are less relevant to qualitative work that is oriented toward hypothesis generation or detailed understanding of the specific participants under study (Reischer & Cowan, 2020). The OSCI might serve as a starting point or model for specialized concept inventories that are better-suited to specific communities of practice that lean to different degrees on quantitative, qualitative, and mixed methods. Our exclusion of some of the original OSCI items following IRT analyses may highlight the tension between reconciling practices across the qualitative and quantitative traditions. For example, responders had a hard time with an item that concerned whether it was appropriate to give corrective feedback to a research assistant during the process of establishing reliability for the qualitative coding of a behavior (identifying pointing gestures in video recordings). The poor fit of this item (item 12) may be due to the lack of consensus (among responders and the scientific community at large) about what the best practices are for training and establishing reliability among coders making qualitative judgments. To move towards a more pluralistic and inclusive open science, researchers must recognize that concepts, such as "reliability," need to be grounded to the different methodological traditions and must develop assessment instruments that reflect the epistemic roots of these traditions.

Related to this point, although for many of the concepts in the OSCI it was straightforward to identify a "best practice," some research dilemmas may have required a more nuanced understanding of the research process, or else more context than what was provided in the vignette. One example of this was "salami slicing" or dividing a project into "minimal publishable units." On the one hand, salami slicing could contribute to a lack of transparency (e.g., if researchers fail to report measures or analyses related to a project that are published in a different paper), and, to a cynical reader, be a product of perverse incentives to maximize the yield of publications. On the other hand, researchers who work with large, complex datasets might pursue multiple research questions that are best addressed in different papers and where it may be impractical to provide exhaustive documentation. Journal editors have acknowledged the difficulty of distinguishing between the legitimate division of large datasets and inappropriate salami slicing (e.g., Werner, 2021). Other questionable research practices may be ethically ambiguous, in that their defensibility depends on the researcher's intention (Sacco & Brown, 2019) or positionality in academia's hierarchical structure. Some of the items that we excluded from the OSCI based on the IRT analysis concerned such arguably ambiguous cases, such as whether to share a developmental dataset that required a lot of effort and resources to collect (item 27), whether to pay an open access publishing fee to make a paper accessible to readers (item 26), and the described example of giving corrective feedback to establish inter-rater reliability (item 12). These excluded items and the low-performing items on the OSCI (e.g., on "salami-slicing") underscore that, for some research dilemmas, readers might require more context to determine the "best" course of action.

A related limitation of the scope of this work is that we focused on undergraduates when developing the OSCI and our instructional materials, and as a result, focused on a core set of concepts that did not depend on advanced statistical training or research experience. Although the OSCI was developed using undergraduate participants, it may nonetheless be a useful pedagogical tool for psychology graduate students, both for assessing misconceptions and for prompting discussions about best practices. As we noted in the introduction, more nuance would be appropriate at more advanced levels of research methods training to convey the different perspectives about methodological reform, theory development, and scientific inference. A further question for future work is how statistical literacy contributes to performance on the OSCI. In our studies we used past coursework as a proxy for statistics knowledge (Study 2) and delivered the survey after students had reviewed key statistics concepts in their research methods course (Study 3). However, we did not independently assess their understanding of statistics concepts such as significance testing, type I and type II errors, and power, which might have figured in their reasoning about specific items on the OSCI. While our initial evaluation of the items in Study 2 assumed that participants' ability varied on a single latent dimension, further work could use multidimensional IRT to explore whether items differ in their reliance on domain-specific knowledge about statistics or more general abilities related to reasoning or reading comprehension.

The implementation of the OSCI in our research methods courses in Study 3 had some additional limitations. First, because data collection was simultaneous in Studies 2 and 3, we were not able to test a more refined version of the instrument (based on the results of IRT in Study 2) in a larger confirmatory sample. Instead, students in Study 3 completed all 40 items but their performance was assessed using the retained set of 33 items of the final OSCI. Second, in our exploratory analysis of performance in Study 3, we excluded a sizable proportion of participants who appeared

to give low effort responses, which reduced our small sample size further. The low effort of some of the students may be unsurprising in the context of the unfolding COVID pandemic during Spring 2020. A related reason for the low effort could be fatigue, whether due to the pandemic or the length of the instrument. Participants may have experienced mental fatigue seeing that the vignettes involved detailed descriptions and that it wasn't always clear what the "best response" was. A reassuring note regarding future work is that, as we noted, the final inventory is shorter, involving 33 items.

The information we provide in this paper can be used to guide the process of shortening the inventory further for future use. The IRT results from Study 2 can guide the selection of items with high discriminability, while taking their difficulty into account. Additionally, the topic description of items (Appendix A) can be used to sample a range of concepts, seeing that in the current instrument some concepts (e.g., publication bias, preregistration, p-hacking) are addressed by two or more items.

Another direction in which the OSCI could be revised for future use concerns its cultural sensitivity. Some of the terms used in the vignettes (e.g., "zip code" in item 23 or "small liberal arts college" in item 27) are specific to the U.S. American cultural context, whose meaning and associations may not be known to readers outside the U.S. Given that institutional structures of higher education and linguistic terms vary across cultural contexts (even among English users), we recognize that the OSCI could be improved to be more culturally sensitive and have broader applicability. In our OSF repository, we include a version of the instrument in which we substitute some of these terms with more neutral or more broadly applicable variants (e.g., "postal code" for "zip code"; see: https://osf.io/8g62m). We encourage future users of the OSCI to review the vignettes for their applicability to the cultural context they're being deployed and tailor them to that context, as necessary.

A methodological consideration of the current work is the lack of a control group in Study 3, which leaves some doubt over whether learning gains among the remaining students were specifically tied to the open science instructional materials. It's possible that other aspects of the course or other changes among students during the course of the semester could have accounted for the observed results. It is unknown how this group would have fared compared to a control group of students that were enrolled in the same course but didn't complete the open science module. Similarly, it's unclear to what extent the new materials contributed to the improvements in self-efficacy beyond the more standard research methods coursework. Our findings serve as a proof of concept for using the OSCI as an assessment, but further work is necessary to establish the pedagogical benefits of our materials, preferably with the involvement of additional instructors, a larger number of students, and additional procedures to better identify low-effort participation such as attention checks. In future work, it would also be beneficial to incorporate positive controls of the instruction materials during the intervention, such as immediate tests on concepts to verify that students attended to the curriculum. Such positive controls would help disentangle the sensitivity of the OSCI from the effectiveness of the instruction material, and would allow for a more meaningful interpretation for any effects (or lack thereof) between pre-test and post-test detected by the OSCI.

Finally, we might also see greater learning gains with a closer integration between the new materials and other class activities. In Study 3, the new materials were delivered independently of other elements of the course. Video lectures were completed through online modules and were not discussed in the context of other class activities in order to create a consistent treatment across sections. This separation also extended to the hands-on lab activities: although the statistics labs were framed around the goal of analytic reproducibility (with students using open datasets to reproduce the results of published papers), there was little explicit connection to the goals or motivation of open science as covered in the video lectures. Standardizing the experience of students from different sections as much as possible, by not linking different elements of the course, served its purpose in our study. However, we expect that closer integration of the new materials with other elements of the course would have the greatest pedagogical benefits.

Conclusions

The field of psychology has seen a rapid transformation in our understanding of the best practices for conducting research over just a few years. Conceptual understanding of these open and transparent research practices – and the problems they are trying to address – will help psychology students reason about scientific evidence, whether as consumers or future producers of research. Toward that end, integrating open science practices into the curriculum and assessing the impact of these efforts is essential. With the OSCI, educators can evaluate conceptual understanding of topics surrounding open science. The OSCI holds promissory potential as it can be used in future research, along with the targeted curriculum we designed, to assess learning gains in this domain with larger samples of students, interventions that involve a more seamless integration between the new curriculum and other elements of the course, and in more advanced courses in which the field's evolving best practices are presented with appropriate nuance. We hope that the open materials we share will contribute to scalable efforts to improve and evaluate students' conceptual foundations for conducting robust and transparent research.

Contributions

Contributed to conception and design: DBM, AG
Contributed to acquisition of data: DBM, AG
Contributed to analysis and interpretation of data: DBM, AG
Drafted and/or revised the article: DBM, AG

Approved the submitted version for publication: DBM, AG

Acknowledgements

We are thankful to Mitra Mostafavi for assistance with the development of vignettes, and to Marviene Fulton and Shaina Glass for assistance with data collection. We are also thankful to Drs. Elise Demeter and Karen Singer-Freeman from the Office of Assessment and Accreditation at UNC Charlotte who provided further input on the vignettes and the survey design. Finally, we thank Drs. Sarahanne Field, Katie Corker, Jordan Wagge, and another anonymous reviewer for suggestions that significantly improved this article.

Funding information

This work was supported in part by funds provided by the University of North Carolina at Charlotte through a Scholarship of Teaching and Learning Program (SoTL) grant and a Scholarship of Assessment grant.

Competing interests

The authors have no conflicts of interest to declare.

Data accessibility statement

All data, materials, and analysis files can be found at the project OSF page: https://osf.io/fejcn. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the studies in this paper.

Our preregistration can be found at https://osf.io/qbtyg. Study 1 was not preregistered due to its qualitative, exploratory nature. Studies 2 and 3 were preregistered before examining any data. All departures from the preregistration are described the main text.

Submitted: March 06, 2023 PDT, Accepted: April 13, 2023 PDT

References

Blincoe, S., & Buchert, S. (2020). Research Preregistration as a Teaching and Learning Tool in Undergraduate Psychology Courses. *Psychology Learning & Teaching*, *19*(1), 107–115. https://doi.org/10.1177/1475725719875844

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and Whether) to Teach Undergraduates About the Replication Crisis in Psychological Science. *Teaching of Psychology*, *45*(2), 158–163. https://doi.org/10.1177/0098628318762900

Ciarocco, N. J., Lewandowski, G. W., Jr, & Van Volkom, M. (2013). The impact of a multifaceted approach to teaching research methods on students' attitudes. *Teaching of Psychology*, *40*(1), 20–25. https://doi.org/10.1177/0098628312465859

Collaborative Open-science Research. (2022). Replications and extensions of classic findings in Social Psychology and Judgment and Decision Making. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/5Z4A8

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805. https://doi.org/10.1098/rsos.200805

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Field, S. M., van Ravenzwaaij, D., Pittelkow, M.-M., Hoek, J. M., & Derksen, M. (2021). *Qualitative Open Science – Pain Points and Perspectives*. https://doi.org/10.31219/osf.io/e3cq4

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education*, *32*(2), 77–96. https://doi.org/10.1080/08957347.2019.1577243

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science*, *7*(6), 600–604. https://doi.org/10.1177/1745691612460686

Gernsbacher, M. A. (2018a). Three ways to make replication mainstream. *Behavioral and Brain Sciences*, *41*, 129. https://doi.org/10.1017/s0140525x1800064x

Gernsbacher, M. A. (2018b). Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability. *Advances in Methods and Practices in Psychological Science*, *1*(3), 403–414. https://doi.org/10.1177/2515245918754485

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Holmes, J. D., & Beins, B. C. (2009). Psychology is a science: At least some students think so. *Teaching of Psychology*, *36*(1), 5–11. https://doi.org/10.1080/00986280802529 35

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jarke, H., Jakob, L., Bojanić, L., Garcia-Garzon, E., Mareva, S., Mutak, A., & Gjorgjiovska, J. (2022). Registered report: How open do you want your science? An international investigation into knowledge and attitudes of psychology students. *PLOS ONE*, *17*(2), e0261260. https://doi.org/10.1371/journal.pone.0261260

Jekel, M., Fiedler, S., Allstadt Torras, R., Mischkowski, D., Dorrough, A. R., & Glöckner, A. (2020). How to Teach Open Science Principles in the Undergraduate Curriculum—The Hagen Cumulative Science Project. *Psychology Learning & Teaching*, *19*(1), 91–106. https://doi.org/10.1177/1475725719868149

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

Kardash, C. M. (2000). Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology*, *92*(1), 191–201. https://doi.org/10.1037/0022-0663.92.1.191

Kelly, C. D. (2019). Rate and success of study replication in ecology and evolution. *PeerJ*, *7*, e7654. https://doi.org/10.7717/peerj.7654

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

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A Practical Guide for Transparency in Psychological Science. *Collabra: Psychology*, *4*(1), 20. https://doi.org/10.1525/collabra.158

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package 'emmeans.'*

Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*(6222), 1261757. https://doi.org/10.1126/science.1261757

Lorenz, T. K., & Holland, K. J. (2020). Response to Sakaluk (2020): Let's get serious about including qualitative researchers in the open science conversation. *Archives of Sexual Behavior*, *49*(8), 2761–2763. https://doi.org/10.1007/s10508-020-01851-3

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pownall, M., Azevedo, F., König, L. M., Slack, H. R., Evans, T. R., Flack, Z., Grinschgl, S., Elsherif, M. M., Gilligan-Lee, K. A., Oliveira, C. M., Gjoneska, B., Kanadadze, T., Button, K. S., Ashcroft-Jones, S., Terry, J., Albayrak-Aydemir, N., Dechterenko, F., Alzahawi, S., Baker, B. J., … Sadhwani, S. (2022). The impact of open and reproducible scholarship on students' scientific literacy, engagement, and attitudes towards science: A review and synthesis of the evidence. *MetaArXiv*. https://doi.org/10.31222/osf.io/9e526

Pratt, M. G., Kaplan, S., & Whittington, R. (2020). Editorial essay: The tumult over transparency: Decoupling transparency from replication in establishing trustworthy qualitative research. *Administrative Science Quarterly*, *65*(1), 1–19. https://doi.org/10.1177/0001839219887663

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology*, *1*(1), 19–38. https://doi.org/10.1146/annurev-criminol-032317-091849

Reischer, H. N., & Cowan, H. R. (2020). Quantity Over Quality? Reproducible Psychological Science from a Mixed Methods Perspective. *Collabra: Psychology*, *6*(1), 26. https://doi.org/10.1525/collabra.284

Rizopoulos, D. (2006). **ltm**: An*R*Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Sacco, D. F., & Brown, M. (2019). Assessing the Efficacy of a Training Intervention to Reduce Acceptance of Questionable Research Practices in Psychology Graduate Students. *Journal of Empirical Research on Human Research Ethics*, *14*(3), 209–218. https://doi.org/10.1177/1556264619840525

Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E.-J. (2020). Teaching Good Research Practices: Protocol of a Research Master Course. *Psychology Learning & Teaching*, *19*(1), 46–59. https://doi.org/10.1177/1475725719858807

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Sizemore, O. J., & Lewandowski, G. W., Jr. (2009). Learning might not equal liking: Research methods course changes knowledge but not attitudes. *Teaching of Psychology*, *36*(2), 90–95. https://doi.org/10.1080/00986280902739727

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). Psychology Press. https://doi.org/10.4324/9781315173726-14

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of Early Adolescence*, *34*(1), 120–151. https://doi.org/10.1177/0272431613511332

van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840x.2020.1853477

Veilleux, J. C., & Chapman, K. M. (2017). Development of a Research Methods and Statistics Concept Inventory. *Teaching of Psychology*, *44*(3), 203–211. https://doi.org/10.1177/0098628317711287

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*, 247. https://doi.org/10.3389/fpsyg.2019.00247

Werner, M. U. (2021). Salami-slicing and duplicate publication: Gatekeepers challenges. *Scandinavian Journal of Pain*, *21*(2), 209–211. https://doi.org/10.1515/sjpain-2020-0181

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Zimmer, F., Draxler, C., & Debelak, R. (2022). Power analysis for the Wald, LR, score, and gradient tests in a marginal maximum likelihood framework: Applications in IRT. *Psychometrika*, 1–50. https://doi.org/10.1007/s11336-022-09883-5

Appendices

Appendix A. Lists of items and topics used in Studies 1 and 2

Table A.1 summarizes the topics associated with the 41 items used in Study 1 to elicit open-ended responses from participants. Items were matched according to topic, to the extent possible, across two lists (A and B). In Study 1, participants were randomly assigned to one of the two lists (A or B). On the basis of coding from Study 1, item 30 was excluded from Study 2, due to difficulties in generating contrasting correct and incorrect options.

For Study 2, the remaining 40 items were adapted to multiple choice questions. Participants responded to all 40 items. On the basis of IRT analysis, items 11, 12, 26, 27, 32, and 34 were excluded from the final version of the instrument; item 13 was also excluded due to being paired with the scenario of item 12. The retained 33 items were used for analyses in Study 3.

Table A.1

	List A	Topic & Concepts	List B	Topic & Concepts
1	item1	Publication bias/file drawer problem: researcher's perspective	item2	Publication bias/file drawer problem: editor's / reviewer's perspective
2	Item3	Preregistration (reader's perspective: evaluating evaluating from preregistered vs. non-preregistered studies)	item34**	Preregistration (researcher's perspective: revising a preregistered plan) / Transparency (about departures from plan)
3	item20	p-hacking: selective reporting of conditions; Transparency (lack of)	item4	p-hacking: optional stopping
4	Item5	Replication (as it relates to Type II errors, statistical power, small sample size, statistical significance)	item7	p-hacking: covariate inclusion; Analytic flexibility; Researcher degrees of freedom
5	Item6	p-hacking: outlier exclusion	item8	p-hacking: dropping participants or observations; Analytic flexibility; Researcher degrees of freedom
6	item17	p-hacking: outcome switching; Transparency (lack of)	item19	p-hacking: selective reporting of outcomes; Transparency (lack of)
7	item22	p-hacking: selective reporting; publication bias (editor's perspective); Transparency	item21	p-hacking: selective reporting of conditions/ experiments; Transparency (lack of)
8	item18	Computational reproducibility	item16	Open data (sharing raw, de-identified data)
9	item27**	Open data (perspective of owner of costly dataset)	item23	Open data (data requester's perspective): privacy issues
10	item24	Non-HARKing (author is clear about exploratory post-hoc analyses); Transparency	item9	HARKing
11	item28	Misconduct: Reusing own data without citation; Transparency (lack of)	item29	Misconduct: Not providing proper attribution for use of open data; Transparency (lack of)
12	item41	Retraction (editor's perspective); Misconduct; Fraud; Transparency	item40	Misconduct; Fraud; Data Fabrication (co-author's / whistleblower's perspective)
13	item30*	Incentive structures: Learning statistical programming / Time costs of open science practices	item26**	Incentive structures: Publishing open access
14	item32**	Incentive structures: Funders (covering open access fees)	item35	Salami-slicing; Incentive structures: Publish or Perish
15	item36	Incentive structures: Publish or Perish / dataset reuse	item31	Preregistration: Time costs associated with open science practices
16	item37	Replication (direct replication vs. conceptual replication)	item39	Replication (statistical power; replicating study using a larger sample size)
17	item33	Data peeking with pilot studies; p-hacking; publication bias; sample size	item38	Replication (perspective of researcher evaluating evidence from a failed replication)
18	item12**	Interrater reliability (establishing interrater reliability)	item14	Misconduct; Fraud; Data Fabrication
19	item13**	Interrater reliability (reporting); Transparency	item15	Misconduct; Fraud; Data Fabrication (reporting fabricated results)
20	Item10	Participant bias (participants not naïve to hypothesis); non-random sampling	item25	Preregistration (Outlier removal based on plan)
21	Item11**	Participant bias (reporting)		

*Note. Item with * was excluded from Study 2 based on coding of responses in Study 1. Items with ** were excluded from analysis in Study 3, based on IRT analyses in Study 2. The remaining 33 items are those selected for use for the OSCI.*

Appendix B. Description of the content of the video lectures developed for the new curriculum

Video lecture 1 (approximately 21 mins). In the first video, we introduce the credibility and replicability crisis in psychology, contextualizing the cultural norms and incentives in science that fuel problematic practices, which distort the scientific record. We describe the scientific pipeline – from the identification of research questions, specification of hypothesis, designing the study, data collection, data analysis, writing, and publication – to illustrate the stages at which problematic practices can occur. We focus specifically on problematic issues at the writing and publication stage (publication bias and the file drawer effect), and on outright unethical cases of research misconduct (e.g., data fabrication) at the data collection and analysis stages. Finally, we introduce retraction as a process of correcting the scientific record.

Video lecture 2 (approximately 22 mins). In the second lecture, we start with the point that while cases of research misconduct get a lot of attention, they're relatively rare. We introduce Questionable research practices (QRP), as conventional behaviors that undermine the robustness of research, linked to the credibility crisis. We describe how biases toward novel findings and significant results influences decisions made at the stages of specifying hypotheses, designing studies, collecting data, analyzing data, drawing conclusions, and reporting research. In the context of an example, we include a refresher of hypothesis testing, and the concepts of false positives and false negatives. We cover practices associated with p-hacking (including optional stopping, dropping observations, outcome switching, covariate inclusion), selective reporting, and hypothesizing after the results are known (HARKing).

Video lecture 3 (approximately 27 mins). In the final lecture, we present some solutions to QRPs and more egregious research practices. We introduce open science practices aimed to correct the systematic failings of scientific practices that promote accessibility, transparency, reproducibility, and robustness in research. We cover the practices of sharing datasets, analysis code, study materials, reporting methods and results transparently and completely, preregistering hypotheses and methods, and planning of sample sizes for research studies. We provide an example of preregistration of each step of the research pipeline: how researchers must specify prior to conducting the study their hypotheses, operational definitions, sampling plan, study design, and analysis plan. We address the distinction between confirmatory and exploratory analyses, and address misconceptions about preregistration. We present resources for implementing these practices (e.g., where to complete preregistrations, how to share data, code, and material). We also address how open science practices facilitate large-scale collaborations. Finally, we cover how these practices are reinforced by other stakeholders in science (including journals, funding agencies, and scientific societies) and how the incentives, requirements, and rewards for open and transparent research have changed (e.g., badges, registered reports, open access publications and preprints).

Appendix C

Pre-test, Post-test, and Difference scores (means and standard deviations) on the proportion correct (Study 3)

Item	Topic	Pre	Post	Post-Pre
25	P-hacking (removing outliers)	0.2 (0.41)	0.52 (0.51)	0.32 (0.69)
2	Publication bias/file drawer problem	0.44 (0.51)	0.72 (0.46)	0.28 (0.61)
37	Replication	0.48 (0.51)	0.76 (0.44)	0.28 (0.68)
8	Data exclusion	0.16 (0.37)	0.4 (0.5)	0.24 (0.6)
33	Data peeking	0.52 (0.51)	0.76 (0.44)	0.24 (0.72)
3	Preregistration	0.6 (0.5)	0.8 (0.41)	0.2 (0.71)
24	HARKing	0.6 (0.5)	0.8 (0.41)	0.2 (0.58)
9	HARKing	0.56 (0.51)	0.72 (0.46)	0.16 (0.47)
18	Reproducibility	0.44 (0.51)	0.6 (0.5)	0.16 (0.69)
28	Misconduct (data citation)	0.28 (0.46)	0.44 (0.51)	0.16 (0.75)
31	Preregistration	0.64 (0.49)	0.8 (0.41)	0.16 (0.55)
23	Open data	0.32 (0.48)	0.44 (0.51)	0.12 (0.53)
1	Publication bias/file drawer problem	0.76 (0.44)	0.84 (0.37)	0.08 (0.49)
16	Open data	0.68 (0.48)	0.76 (0.44)	0.08 (0.49)
21	Selective reporting (conditions)	0.76 (0.44)	0.84 (0.37)	0.08 (0.49)
4	Optional stopping	0.8 (0.41)	0.84 (0.37)	0.04 (0.61)
5	Replication	0.68 (0.48)	0.72 (0.46)	0.04 (0.54)
7	Covariate inclusion	0.6 (0.5)	0.64 (0.49)	0.04 (0.61)
17	Outcome switching	0.8 (0.41)	0.84 (0.37)	0.04 (0.35)
38	Replication	0.72 (0.46)	0.76 (0.44)	0.04 (0.61)
39	Power/sample	0.24 (0.44)	0.28 (0.46)	0.04 (0.54)
10	Participant bias	0.92 (0.28)	0.92 (0.28)	0 (0.41)
19	Selecting reporting (outcomes)	0.72 (0.46)	0.72 (0.46)	0 (0.71)
22	Selecting reporting (editor's perspective)	0.76 (0.44)	0.76 (0.44)	0 (0.58)
29	Misconduct (data citation)	0.84 (0.37)	0.84 (0.37)	0 (0.5)
41	Retraction	0.76 (0.44)	0.76 (0.44)	0 (0.58)
14	Misconduct (data fabrication)	0.8 (0.41)	0.76 (0.44)	-0.04 (0.54)
35	Salami slicing	0.4 (0.5)	0.36 (0.49)	-0.04 (0.73)
36	Incentive structures	0.64 (0.49)	0.6 (0.5)	-0.04 (0.45)
15	Misconduct	0.88 (0.33)	0.8 (0.41)	-0.08 (0.49)
20	Selective reporting (conditions)	0.8 (0.41)	0.72 (0.46)	-0.08 (0.49)
40	Misconduct (data fabrication)	0.72 (0.46)	0.64 (0.49)	-0.08 (0.57)
6	P-hacking (data exclusion)	0.76 (0.44)	0.6 (0.5)	-0.16 (0.69)

Appendix D

Descriptive statistics for Pre-test, Post-test, and Difference scores (means and standard deviations) on the Questionnaire on Attitudes toward Research (N = 25) in Study 3

Item	Pre	Post	Post-Pre
General attitudes			
1. I like psychology	4.96 (0.2)	4.96 (0.2)	0 (0)
2. I think psychological research is similar to research in fields like chemistry, physics, or biology	3.8 (0.65)	3.96 (0.84)	0.16 (1.11)
3. I think psychological research is similar to research in fields like philosophy, literature, or modern languages	3.56 (1.04)	3.84 (1.25)	0.28 (1.43)
4. I think psychology is a soft science	2.16 (1.07)	2.48 (1.29)	0.32 (1.49)
5. I think most psychology findings are just common sense	1.88 (0.73)	2.32 (1.18)	0.44 (1.29)
Trust			
6. I trust the results of psychology studies that follow the scientific method	4.44 (0.58)	4.32 (0.63)	-0.12 (0.67)
7. I trust the results of psychology studies to be able to be replicated	4.44 (0.71)	4.48 (0.65)	0.04 (0.93)
8. I trust the results of studies in the psychology literature	4.24 (0.6)	4.04 (0.79)	-0.2 (1.04)
9. I trust the results of psychology studies that receive a lot of media attention	3 (0.76)	2.92 (1)	-0.08 (1.35)
10. I trust the results of psychology studies with counterintuitive or surprising results	3.48 (0.82)	3.36 (0.81)	-0.12 (1.05)
All "trust" items	**3.92 (.46)**	**3.82 (.49)**	**-.1 (.65)**
Interest			
11. I am interested in pursuing graduate school in psychology	4.4 (0.87)	4.28 (1.1)	-0.12 (1.36)
12. I am interested in pursuing a career in a field related to psychology	4.68 (0.75)	4.6 (0.87)	-0.08 (1)
13. I am interested in using the results of psychology research in my future career	3.68 (1.22)	3.96 (1.1)	0.28 (1.59)
14. I am interested in conducting psychology research in my future career	3.32 (1.22)	3.4 (1.26)	0.08 (1.63)
All "interest" items	**4.02 (.72)**	**4.06 (.86)**	**.04 (1.02)**
Self-efficacy			
15. I am confident in my ability to concentrate and stay fully focused on the materials being presented throughout each class period.	3.96 (0.73)	4.24 (0.66)	0.28 (0.98)
16. I am confident in my ability to memorize and recall on demand the facts and concepts covered in my classes.	3.88 (0.73)	3.88 (0.6)	0 (1.04)
17. I am confident in my ability to focus exclusively on understanding and answering questions and avoiding breaks in my concentration during exams.	3.56 (1)	3.8 (1.04)	0.24 (1.59)
18. I understand the facts, concepts and arguments covered in my classes as they are presented in lectures and the textbook.	4.16 (0.75)	4.4 (0.5)	0.24 (0.93)
19. I am confident in my ability to explain the facts, concepts and arguments covered in my classes clearly to others in my own words.	3.72 (0.74)	4.24 (0.72)	0.52 (1.08)
20. I am able to discriminate between the more important and less important facts, concepts and arguments covered in my classes.	4.04 (0.84)	4.44 (0.58)	0.4 (0.71)
21. I am able to make understandable course notes which emphasize, clarify and relate key facts, concepts and arguments as they are presented in lectures and the text.	4.32 (0.56)	4.24 (0.72)	-0.08 (0.86)
22. I am confident in my ability to perform appropriate statistical analyses given a research design	3.24 (1.05)	3.96 (0.73)	0.72 (1.06)
23. I am confident in my ability to deliver a presentation about a research project	3.64 (1.08)	4.28 (0.68)	0.64 (1.22)
24. I am confident in my ability to write a research report	3.48 (1.08)	4.28 (0.61)	0.8 (1.19)
All "self-efficacy" items	**3.8 (.57)**	**4.18 (.45)**	**.38 (.70)**

Appendix E

Table 1. Results of linear mixed effects regression models for Study 3.

	Estimate	SE	2.5%	97.5%	*t*	*p*
Estimated ability						
(Intercept)	0.289	0.169	-0.034	0.612	1.711	0.101
Timepoint	0.182	0.074	0.034	0.33	2.452	0.022
Semester	0.149	0.169	-0.174	0.471	0.881	0.388
Instructor	0.051	0.167	-0.268	0.369	0.306	0.763
Self-efficacy						
(Intercept)	3.992	0.088	3.823	4.161	45.204	<0 .001
Timepoint	0.188	0.061	0.065	0.311	3.062	0.005
Semester	0.017	0.088	-0.152	0.186	0.192	0.85
Instructor	-0.061	0.087	-0.228	0.106	-0.701	0.491
Trust						
(Intercept)	3.867	0.088	3.698	4.036	43.827	<0 .001
Timepoint	-0.048	0.047	-0.142	0.046	-1.015	0.32
Semester	0.047	0.088	-0.122	0.216	0.534	0.599
Instructor	-0.038	0.087	-0.204	0.129	-0.433	0.669
Interest						
(Intercept)	4.028	0.154	3.733	4.323	26.071	<0 .001
Timepoint	0.02	0.064	-0.107	0.147	0.314	0.756
Semester	0.003	0.154	-0.292	0.298	0.019	0.985
Instructor	0.096	0.152	-0.195	0.388	0.632	0.534

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/75226-development-of-a-concept-inventory-on-open-and-transparent-research-practices/attachment/158988.docx?auth_token=7CpfRTHDoE5dqsVP5GFB