Replication Crisis: Pre-Post Educational Video Assessing Psychology Students' Behaviors and Attitudes

by

Gillian Foster

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of

The University of Manitoba

In partial fulfillment of the requirement of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

In the past decade, replication, crises surrounding replication, fraud, and data collection have dominated discussions in psychology. Open Science, a cooperative approach using digital technologies, addresses these issues. This study explored Honours (PSYC 4520) and First - Year (PSYC 1200) Undergraduate Psychology students' Attitudes Toward Psychology, the Replication Crisis, and their Engagement in proper Open Science Practices. 168 students participated in a pre-test post-test educational video intervention assessing their attitudes and knowledge (post-test only). An 8-item questionnaire was given to Honours students (n = 13) at the end of the academic year related to the Open Science protocols they engaged in (e.g., pre-registration, power analysis, open-data sharing, etc.) for their honours thesis. A repeated-measures t-test examined changes in within-subject participants' engagement and attitudes pre-to-post intervention, assessing how responses evolved over time. A mixed-design ANOVA was used to investigate the interaction effect between level of study (Honours vs. First-Year) and time points (before and after video) on Engagement in proper Open Science Practices. The intervention led to meaningful changes in students' Attitudes toward Psychology and their Engagement in proper Open Science practices. However, no significant change was found in Attitudes toward the Replication Crisis. Replication Crisis Knowledge was similar across both academic levels, and students who demonstrated greater understanding of replication issues tended to report more willing Engagement in proper Open Science practices. This study has significant implications on the importance of properly practicing Open Science in one's undergraduate degree and gaining proper experience for future study management.

Keywords: replication crisis, open science, attitudes, knowledge, pre-test post-test intervention

Acknowledgements

I would like to express my deepest appreciation to my advisor Dr. Johnson Li for his guidance, invaluable feedback, and encouragement throughout the writing process. I would also like to thank my committee members Dr. Wan Wang, and Dr. Nicholaus Brosowsky for their insights and expertise. Finally, I would like to thank my mother, brothers, sister-in-law, and aunt for the motivation and support during this process.

Table of Contents

Replication Crisis: Pre-Post Educational Video Assessing Psychology Students' Behaviors and Attitudes

In the last 10 years, crises surrounding replication, fraud and best data collection practices, and reporting have governed discussions in the field of psychology (Chopik et al. 2018). "The replication crisis is a term used to refer to widespread concern about the accumulation of erroneous studies in scientific literature" (Forbes et al., 2023, p. 237). Consequently, recent studies published in high-ranking journals have scrutinized in various replication studies: Camerer et al. (2018) found that only 13 out of 21 behavioral and social science studies published between 2010 and 2015 in the journal *Nature and Science* could successfully replicate (Renkewitz & Heene, 2019). Additionally, Open Science Collaboration (2015) wrote that with estimated replication rates between 25% (social psychology) and 50% (cognitive psychology), it was clear that psychology endured from a severe replicability problem.

At the core of the crisis is the increasingly alarming realization that common research practices are indeed problematic, and discussions began to surface about the reasons why replication rates were disappointingly low. Some reasons such as poor study design (i.e., low statistical power; Button et al., Ioannidis, 2005; as seen in Sarafoglou, 2019), the field's unwillingness to conduct direct replication studies (Pashler & Harris, 2012; Schmidt, 2009; as seen in Sarafoglou), "and a bias to selectively report positive results" (Francis, 2013; Scargle, 1999; as seen in Sarafoglou, 2019, p. 47) were known for decades, but largely ignored (Sterling, Rosenbaum, & Weinkam, 1995). Furthermore, several researchers admit to questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012; as seen in Sarafoglou, 2019), but were never systematically investigated (John, Loewenstein & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; as seen in Renkewitz & Heene, 2019). QRPs are activities that are not transparent, ethical, or fair, and thus threaten academic integrity and the publishing process. They

are hard to identify and define, and sometimes easy to get away with, which is what makes them "questionable" rather than technically illegal. QRPs and research misbehaviors are decisions made during the research process that raise questions about your work's rigor and precision. Some examples include *p*-hacking, degrees of freedom "*calibrations"*, variance manipulation, and sample manipulation (Suter, 2020). From poor data collection to QRPs, this encompasses just a fraction of what I call the replication crisis.

The "Real" Replication Crisis

Difficulties Powering a Replication Study

Psychology is not the only natural science where this occurs; the replication crisis extends to other fields such as economics (Camerer et al., 2016; as seen in Beaudry, 2022), and medicine (Errington et al., 2021; as seen in Beaudry, 2022). Meanwhile, I am solely focusing on psychology and there are difficulties in adequately conducting a replication study. First, a decision must be made as to how much power is in fact adequate. This can be done using statistical software called G-power, where you will input a desired power and estimated effect size to determine the sample you will need to reach the power estimate. According to Cohen (1988), it is more standard to design a study to have statistical power of .80. This means that there is an 80% chance of rejecting the null hypothesis if it is false (i.e., finding a true effect). Although choosing a statistical power value can be subjective, only leaving a 20% chance of not finding a true effect does not seem responsible. Maxwell et al. (2015) speak about a real-world example that is pertinent to replicable studies. To summarize, let us suppose 100 published studies correctly rejected 100 different false null hypotheses. Should I be satisfied that if I attempted to replicate each of these studies, only 20 of these findings cannot be trusted? There is a strong argument that replication studies should have a higher statistical power than .80 such as

.90 or .95, therefore this will reduce the number of non-trusted studies and increase the likelihood of a significant test detecting an effect when there is one.

Second, even after identifying your desired power value, the appropriate sample size depends on the presumed effect size. An effect size quantifies the magnitude or strength of the difference between groups or the relationship between variables. In fact, designing a replication study can be advantageous compared to designing an original study because the effect size value being available from that original study (Maxwell et al., 2015). However, there are times where that is not the case. An immediate problem is that, in the literature, researchers generally make biased estimates of the true population effect sizes. Examples of this can be publication bias and selective reporting. Publication bias appears when a study that is published in a peer-reviewed journal is published based on the strength of the results rather than individual factors (Nikolopoulou, 2023). Additionally, selective reporting bias can have serious consequences. If data from the study is excluded, the 'true effect estimate' of the results can be over or underestimated, therefore misleading the results (Baur, 2022). Publication bias and self-reporting bias can lead to published effect size values that are larger than their population counterparts (Greenwald, Gonzalez, Harris & Guthrie, 1996; Lane & Dunlap, 1978; Maxwell, 2004; Schmidt, 1992; as seen in Maxwell et al., 2015). Hence, a replication study that is designed to have adequate power can be underestimated and will more likely fail to replicate the original study.

Third, when a replication study bases their sample size on the effect size of the original study, researchers fail to account the sampling variability in the original sample effect size. Statisticians have two separate occasions of this occurrence called "conditional power" and "predictive power". Conditional power is the probability of rejecting the null hypothesis dependent on a presumed effect size that is supposed to be known with confidence. On the other

hand, predictive power acknowledges that the effect size is not known but rather considers it as an estimate. It averages the power over the possible values of effect size. This is generally based on the estimated standard error of the estimated effect size. By doing so, this will obtain a point estimate of the power while considering uncertainty (Maxwell et al., 2015). According to Dallow and Fina (2011), predictive power can lead to larger samples sizes compared to conditional power when it is used with the same value for power. Researchers and psychologists that use power analysis to design a replication study generally use conditional power calculations. This means they implicitly assume a single value of the unknown population effect size (Maxwell et al., 2015). However, this can bring about undersized, underpowered studies (Dallow & Fina, 2011). So, even if published effect sizes are not biased, it is unlikely for replication studies to control Type II error rates at the desired level.

Maxwell et al. (2015) described a hypothetical original study to further illustrate the importance of sampling variability. Let us say a study compared the means of two independent groups with 40 participants per group (i.e., 80 participants in total) and produced a statistically significant *t* value of 2.24. A researcher decides they want to replicate this study, and asks themselves how large does the sample size have to be for conditional power to equal the .80 conventional statistical power standard? The *t* value of 2.24 corresponds to a 0.50 (“medium” effect size) Cohen’s *d* value (Cohen, 1988), which implies – if ran through *G*-power – those 64 participants (128 in total) per group is needed to achieve a power of .80. However, the effect size of 0.50 is only an estimate. The true population effect size could be much smaller or much larger than 0.50. For example, when $n = 40$ per group, there is a 50% confidence interval that the effect size ranges from 0.35 to 0.65. Therefore, there is a 25% chance that the population value is less than .35, just as there is a 25% chance that the population value is more than .65. How powerful

will the replication study with a sample size of 64 per group be if the true effect size is only 0.35 or is as large as or larger than 0.65? Let us say the true effect size is 0.65. If I compute the achieved power in *G*-power, it corresponds to .95 power, which is larger than the anticipated power of .80. One could expect a similar drop in power if it is smaller than .50. For a two-tailed independent group means where the true effect size is 0.35, the power drops down to .50. Based on this hypothetical study, there is a 25% chance that the true power is no more than .50. In fact, the statistical power would be much less, around .35 if the replication study is designed with the same sample size as the original study (i.e., 40 per group).

The rational implication here is that basing sample size calculations on the effect size from the original study is not as wonderful as it may seem. A researcher may feel confident about having a power of .80 for their replication study, but they may be confronted with a much lower power. What could be contributing here is the nonlinear relationship between effect size and power. From the previous example, a statistical power of .80 corresponds to an effect of 0.50. Even though the effect size of 0.50 is in the middle of 0.35 and 0.65, their statistical power values of .50 and .95 are not. The statistical power of .80 for an effect size of 0.50 is not equidistant from .50 and .95. In fact, in the world of research, it is a much greater loss to have a low effect size-low power, compared to the power gain with a high effect size. As a result, not taking sampling variability into account while designing a replication study can have a greater risk for an underpowered study. Despite the rigorous effort required to power a replication study adequately, even well-designed replications can yield non-significant results. This outcome introduces a new layer of complexity: interpreting what a failure to reject the null hypothesis means. The challenges of powering a study – such as biased effect size estimates, sampling variability, and reliance on conditional power – do not just affect the design phase; they also cast

doubt on the conclusions drawn from replication attempts. In other words, the difficulty does not end with calculating sample sizes – it extends into the murky territory of interpreting what a non-significant result truly implies about the original finding. This leads us to the next major hurdle in replication science: the difficulty of interpreting non-significant results.

Difficulties Interpreting Non-Significant Results

The previous section focused on why designing a replication study to have adequate power is not always that simple. Now, let us suppose it is possible to design a replication study with adequate power, but it fails to produce a statistically significant result. At this point, it would be common for researchers to stop their research due to the replication "failing", and this can be due to the original study's results being overturned. This section will focus on why this interpretation is hasty. Maxwell et al. (2015) explains a relevant example like the previous one. Let us consider the means of two independent groups of 40 people ($N = 80$) were compared in an original study, and they reported a statistically significant *t*-value of 2.24. The researcher wants to replicate this study and starts by calculating the effect size from the original study. Cohen's *d* is 0.50. By following conventional practices at an alpha level of .05 two-tailed test, the researcher chooses a sample size of 86 per group – 172 in total – for the replication study, because this provides a statistical power of .90 to detect a medium effect (Maxwell et al., 2015). Notice here that a power of .90 needs more than twice as many participants per group as the original study.

Next, let us suppose the replication study produces a *t*-value of 1.50 and the subsequent two-tailed *p*-value is .14, therefore this replication study does not have a statistically significant result (Maxwell et al., 2015). Conventional practices would conclude that the replication study failed, and the original study did not present strong validity. Particularly, in the original study, if

the sample effect size is the population effect size, the fact that having a sample of 86 per group with a power of .90 guarantees there is only a 10% chance of not obtaining a statistically significant result in your replication study. Some participants in the replication debate might conclude the replication fails to support the original study, and the original study's results should be taken with precaution. Furthermore, some individuals will mistakenly infer that the null hypothesis is likely true. What does this mean? In simpler terms, the effect size is mistakenly interpreted as zero (or almost zero). These are often the sorts of conclusions when replication studies produce non-significant results.

Non-significant results hold a very important question, to what degree do nonsignificant results support the legitimacy of the null hypothesis? To answer this question, I can go over the previous example in more detail. Maxwell et al. (2015) hypothetically described that the *t*-value of 1.50 sample value of Cohen's *d* was 0.23 for the replication study. A 95% confidence interval for the population value of 0.23 is –0.07 to 0.53. Because zero is a part of this interval, it is possible for zero to be the true population effect size. However, it is different to conclude that it is the true population effect size that can possibly be zero rather than the effect size equals exactly zero. The results of the study do not support the conclusion that the effect size is anywhere near close to zero. Considering the confidence interval, the results of the replication study demonstrate that the population effect size could be 0.50. It is ill-suited to say that there is no effect when it is possible there could be a medium size effect (i.e., 0.50).

Although the researcher who followed the replication study did everything according to plan and designed the study very carefully, they cannot confirm nor deny the original study. Although it is possible that the population effect size is zero, it is just as possible for it to be a medium effect size as 0.50 or even larger as seen in the original study. This means that the non-

significant result obtained in the replication study does not hold strong evidence to support that the null hypothesis is true. Even though the replication study could not confirm the original study, that does not mean the replication study should be dismissed. It simply could neither confirm nor deny the results of the original study.

The Issue of Interpreting Null Results

The main issue is how to justify the conclusion that the results from the original study are not trustworthy (Maxwell et al., 2015). The answer seems simple, the nonsignificant results in the replication study justify dismissing the original study if the replication study has adequate power. However, the vital point here is that this simple answer is wrong. The fact that a replication study resulted in nonsignificant results does not mean I should dismiss the original study.

The statistical problem is that to reject the original study's results, I must accept the null hypothesis of the replication study. Let us say I design a replication study where I might interpret that individuals who are exposed to subtle social cues will behave the same way as individuals who are not exposed to those cues, but failing to reject the null hypothesis is not the same as accepting it. This seems to be the norm when teaching students, and what seems to be forgotten nowadays is interpreting those non-significant results in the replication study. In fact, of articles reporting at least one non-significant result, 66.7 % show proof of false negatives (Hartgerink, 2017). The predicament here is how do I use the data from a replication study to decide whether to dismiss the results of the original study in favor of a conclusion where there is the absence of an effect. There is a solution to this problem, frequentist statistics.

Frequentist statistics. Frequentist statistics are about long-run probabilities; the data set collected and analyzed is one of many hypothetical datasets addressing an identical question, and uncertainty is due to sampling error (Fornacon-Wood et al., 2022). For example, the probability of getting tails when flipping a coin in the long run is always 0.5; if I flip the coin many times, I'd expect to see tails 50% of the time, however if I flipped the coin only a few times, I'd expect to observe a different distribution (e.g., all tails) by chance. For a frequentist, they assume the null hypothesis to be true before data collection (e.g., there is no real effect of a particular treatment on survival). To answer this question requires a so-called 'region of equivalence'. A region of equivalence represents a range of parameter values for which the null hypothesis is true (Maxwell et al., 2015).

This idea is related to Serlin and Lapsley's (1985) concept; the "good enough principle" (as seen in Maxwell et al., 2015) to support a theory. This approach requires scientists to establish how close to a specific theoretical value (e.g., zero) an effect needs to be to conclude that the similarity between the data and the theory is good enough that the data supports the underlying theory. Maxwell et al. (2015) provide an example to explain this non-straightforward approach. Let us say researchers decide that a *Cohen's d* value between –0.10 and 0.10 is good enough when the underlying theory predicts a true null effect. However, it is not that simple. It is important to note that this region is in terms of population parameters, rather than sample observations. As a result, it is not good enough to see if the sample value of Cohen's *d* falls within the interval. Rather, it is necessary to consider sampling variability. I can perform a statistical test (Rogers, Howard & Vessey, 1993; as seen in Maxwell et al., 2015) or a confidence interval (Seaman & Serlin, 1998; as seen in Maxwell et al., 2015) to have sampling variability. Because a statistical test takes more time to see a true result, it is more convenient from the

perspective of a confidence interval, and the results of a replication study can fall into one of three categories.

First, a confidence interval of the effect can fall in the equivalence region (Maxwell et al., 2015). Let us say that our confidence interval for Cohen's *d* falls within the –0.10 to 0.10 interval. I would say that the effect is essentially zero at its highest and lowest value, taking sampling variability into consideration, therefore I am highly certain that the true population effect size is no larger than 0.10 or lower than – 0.10. In this case, the conclusion is that the effect size is essentially zero. This result does provide strong statistical evidence but is criticized to not be theoretically or practically significant. It is more statistically significant than the original study that found a significant effect. Second, the confidence interval can fall outside the equivalence region (Maxwell et al., 2015). This suggests that the effect size cannot be close to zero. Not only is the effect statistically significant, but it has implications. A result like this supports an original study that found a statistically significant effect. Third, "the confidence interval can overlap the equivalence region" (Maxwell et al., 2015, p.493). For example, the 95 percent confidence interval for the first group mean could be (5.6, 14.7), and the confidence interval for the second group mean could be (12.6, 19.7). They overlap here, and unfortunately this suggests that the effect may be questionable, but it also might not be. It is ambiguous because the result doesn't contradict or support an original study that found a statistically significant effect.

To understand these possible outcomes, I can consider a hypothetical situation that is as described in Maxwell et al. (2015). Let us say I obtained a *t* value of 1.50 with 86 participants per condition. The 90 percent confidence interval for the population value of Cohen's *d* is an interval of –0.02 to 0.48 which overlaps the equivalence region. As stated before, the results of

the replication study are ambiguous, therefore it is difficult to say whether the replication study accepts or denies the original study. To understand this example and how it relates to hypothesis testing, I can consider a study conducted by Wortman et al. (2014) (as seen in Maxwell et al., 2015). These authors aimed to replicate a study by Bargh and Shalev (2012), where the experience of physical warmth leads to reduced feelings of loneliness. Bargh and Shalev (2012) obtained a Cohen's *d* of 0.61, while Wortman et al. (2014) obtained a Cohen's *d* of 0.02, which is almost zero. At first, it seems like I should reject the original study if the replication had enough power. Wortman et al. (2014) went through extremes to ensure their study was adequately powered. They ended up with 260 participants while the original study had 75 participants. Furthermore, the 90% confidence interval for Cohen's *d* of the replication study ranged from –0.18 to 0.22. This means that, based on their data, they can be 90% confident that the true population effect size lies within this range. Interestingly, this range includes values up to 0.22, which is larger than Cohen's benchmark for a "small" effect size (0.20). Therefore, it is possible that the true effect size is non-zero, even though the replication study found a Cohen's *d* value of almost exactly zero. In other words, even if a replication study that is adequately powered (i.e., has a large enough sample size to detect an effect if one exists) finds no effect, this does not necessarily mean that the effect does not exist. It could simply mean that the effect is smaller than what the original study suggested. This highlights the importance of considering effect sizes and confidence intervals, not just whether an effect is statistically significant.

Potential Solutions

While frequentist statistics can help figure out dismissal of the original study, it can be misleading. That is why there are other solutions that can help address this replication crisis such as conducting more replications, larger sample sizes, thoroughly tested measures, preregistration,

transparent replications, and open science practices. Replication is the process of repeating a study to see if its findings can be independently reproduced (Metskas, 2022). This is a fundamental aspect of scientific research, as it helps to ensure that the results are not just a one-time occurrence but can be consistently observed. However, replication studies are often seen as less prestigious and are therefore less likely to be conducted. Next, larger sample sizes increase the statistical power of a study, making it more likely to detect a true effect if one exists. However, larger sample sizes require more resources, which can be a barrier for many researchers. However, larger sample sizes require more resources, which can be a barrier for many researchers (Metskas, 2022). Furthermore, using thorough test measures can help ensure that the methods used in a study are valid and reliable. This means that the measures accurately reflect the construct they are intended to measure and produce consistent results over time.

Furthermore, preregistration involves specifying the study's hypotheses, methods, and analyses before the data are collected (Science, n.d.). This can help prevent practices such as p-hacking or data dredging, where researchers test multiple hypotheses and only report those that yield significant results. In addition, the Transparent Replications project aims to replicate a substantial fraction of papers published in top psychology journals (*» About*, 2024). The project rates papers on their transparency, replicability, and clarity, which can incentivize researchers and journals to prioritize these qualities. Finally, I have one of the most prominent strategies to help with the replication crisis, Open Science.

Open Science. Open science is a broad term that represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools. For example, *Psychological Science* (the flagship journal of the Association for Psychological Science) and other journals now issue

what's called 'digital badges' to researchers who pre-registered their hypotheses and data analysis plans, openly shared their research materials with other researchers (e.g., to enable attempts at replication), or made their raw data available to other researchers (see Figure 1; as seen in Jhangiani & Chiang, 2015). These initiatives, thanks to Center for Open Science, lead to the development of "Transparency and Openness Promotion guidelines" which has been formally adapted by more than 500 journals and 50 organizations (Jhangiani & Chiang, 2015).

There are three open science practices that should be taught: pre-registration, open data sharing, and open-access publishing (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). Pre-registration is a key open science practice that involves registering a study design and analysis plan publicly before conducting your research (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023).

Pre-registration helps increase transparency and credibility of research by reducing the risk of researchers changing their hypotheses or analysis after the data has been collected, which can ultimately lead to biased or misleading results. Pre-registration can be done through various platforms such as ClinicalTrials.gov, AsPredicted, or Open Science Framework (OSF), and typically involves providing information about your study such as research question(s), hypotheses, sample size, research design, statistical analysis plan and expected outcomes (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). There are two main types of pre-registration: registered reports and pre-results review (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). To have registered reports, researchers submit a detailed research proposal to a journal before data collection. The journal will send you feedback and a decision on acceptance of the study

based on the quality of your research design and methods. To have pre-results reviewed, researchers can submit a detailed research proposal to a preprint server or repository, and reviewers can provide feedback on the quality of the research design and methods. Pre-registration is particularly important when addressing publication bias, which occurs when studies with positive results are more likely to be published compared to those with negative results or inconclusive results (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). By registering a study design beforehand, researchers can reduce this bias by ensuring their study is evaluated based on quality and accuracy.

Open data sharing is another essential open science practice that involves making your research data openly accessible to other researchers, through online platforms or data repositories (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). By sharing data openly, researchers can promote transparency, collaboration and reproducibility which will encourage researchers to build on their work. On top of reproductibility and collaboration, open data sharing promotes cost savings (e.g., the reuse of existing data) and innovation (e.g., development of new data analysis, technologies, and methods). Several data platforms and repositories support open science, such as Dryad, Figsahre, Open Science Framework, and Zenodo (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). These platforms allow researchers to upload and share their data sets and provide documentation to help other researchers understand and be able to use the data. Although open data sharing has several advantages, it has its challenges, such as ensuring the protection of sensitive or confidential data and ensuring that the data is properly documented and formatted to be useful to other researchers (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). On the other hand,

researchers and organizations are actively working to overcome these challenges and promote open data sharing.

Open-access publishing is another important open science practice that promotes making scientific research articles freely accessible to the public without restrictions (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). Open-access publishing helps address the issues of limited access to scientific literature, and high subscription fees. There are two types of open-access publishing: gold open access and green open access. Gold open access allows the researcher's article to be free by the publisher of the journal, and makes it available online immediately upon publication, and covers the costs of publication through article processing charges (APCs) which are paid either by the author or the institution (e.g., university or company) (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). However, green open access allows the researcher's study to be published in a subscription-based journal but also makes a copy of the articles available in a repository or on their personal website. This step is usually after an embargo period which is between three months to two years (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023).

There are many benefits to open-access publishing including: greater collaboration and reuse, reduced costs, and increased visibility and impact (*Open Science Practices and Methods: What They Are, Why They Are Important, and Examples*, 2023). Open-access publishing supports researchers to collaborate more effectively and allows for reusing and adapting research articles for other purposes. In addition, open-access publishing can help reduce the costs of accessing scientific articles and accessing scientific information. Lastly, it helps make scientific research widely visible and accessible, which can increase impactful research. Journals such as PLOS

ONE, BMC, and Frontiers, and more traditional subscription-based publishers, such as Elsevier, and Springer, offer open-access options.

Integrating Statistical Inference and Open Science Practices

A nuanced grasp of both statistical inference and Open Science Practices is crucial for students to navigate the replication crisis. Interpreting null results – whether through confidence intervals, ROPE-based HDIs, or Bayes factors – demands deep statistical reasoning, while solutions like preregistration, open data, and transparent replications require practical data-management skills and a culture of accountability. Teaching open science must therefore weave these threads together: students need hands-on experience with pre-registering studies, sharing datasets, and critically evaluating replication outcomes alongside traditional hypothesis tests. Embedding real-world case studies and scaffolded exercises into the curriculum will prepare future researchers to both interpret ambiguous findings and uphold transparent methods. Let us turn our attention to a review of previous studies that have embraced these tenets of Open Science, to better understand the real-world impact and challenges of implementing these principles.

Previous Studies

Sarafoglou et al. (2020)

The study titled “Teaching Good Research Practices: Protocol of a Research Master Course” focused on the importance of teaching open science practices to students, particularly in the context of the current crisis of confidence in psychological science. The course described in the study was designed and taught at the University of Amsterdam and aimed to enhance transparency, reproducibility, and replicability in research. The course is supported by Chambers’ (2017) book “The 7 Deadly Sins of Psychology” and covers topics such as QRPs, the

significance of direct and conceptual replications studies, preregistration, and the public sharing of data, code, and analysis plans. A pedagogical approach was adopted that minimized teacher-centered lectures, emphasized practical training in open science practices, and encouraged student engagement in open science discussions on social media platforms. The course included regular classes and classes organized by students, with an example being a class focused on "The Sin of Data Hoarding," discussing the benefits of data sharing and the cultural reluctance to share data in the scientific community. The study underscores the necessity of integrating open science practices into student education to solidify field-wide reforms and build a more reliable and transparent scientific community. In Sage Journals, this study has gold-open access, open-access publishing, open access materials, and was pre-registered.

Beaudry et al. (2022)

The study titled "What do Incoming University Students Believe About Open Science Practices in Psychology" aimed to understand the incoming undergraduate students of psychology's belief about reproducibility and open science practices. Reproducibility and open science practices are fundamental to the scientific method, and understanding students' perceptions of these is crucial for educators. The researchers conducted an online survey with participants who were about to start their first psychology course at a university. University students were answered from Australia, New Zealand, the United Kingdom, and the USA. The total number of participants was 239.

The survey asked the students about their beliefs on how a researcher should conduct their study. The results showed that most students supported several open science practices. Open science practices refer to the idea that scientific research should be transparent, accessible and reproducible. This includes practices such as sharing data, methodologies, and findings openly

(Beaudry et al., 2022). However, when the students were asked to estimate the proportion of published psychological studies that follow these open science practices, the average estimate was around 50%. This suggests that while students support these practices, they may overestimate how commonly they are used in actual research. Interestingly, only 18% of participants reported that they had heard the term “replication crisis”.

The study concluded that despite the media attention about the replication crisis, few incoming psychology students in the sample were familiar with the term. However, the students were generally in favor of most open science practices, although they overestimated the prevalence of some of these practices in psychology. The implications of this study for teaching suggest that teachers of incoming psychology students should not assume pre-existing knowledge about open science or replicability. This highlights the importance of incorporating these topics into the curriculum to ensure students have a comprehensive understanding of current scientific practices. This could potentially lead to more robust and reproducible research in the future as these students become the next generation of researchers. In Sage Journals, this study has green-open access, open-access publishing, open access materials, and was pre-registered.

Chopik et al. (2018)

The study titled “How (and Whether) to Teach Undergraduates About the Replication Crisis in Psychological Science” explored the challenges and methods of educating undergraduate students about the replication crisis in psychology. Specifically, these issues have sparked discussion on how to best communicate and teach these topics to undergraduate students. The study aimed to examine the effects of a one-hour lecture designed to communicate the replication crisis and recommendations to increase reproducibility in psychological research.

194 undergraduate students completed one survey the week prior to the lecture regarding the replication crisis and answered the same survey the week after.

At both time points, students completed attitudinal questions about the field of psychology, read a news article for a study that had generated a lot of attention but had difficulty replicating, and answered questions about the article. The week prior, participants were randomly assigned to read a news article describing either interpersonal effects of physical warmth (Lynott et al., 2014; William & Bargh, 2008) or how a single exposure to the American flag shifts political views 8 months later (Carter, Ferguson, & Hassin, 2011; Klein et al., 2014). The week after, participants read the other news article not previously read and were asked the same questions.

As a result, the surveys indicated that the lecture was an effective pedagogical tool. After the lecture, students showed a slight decrease in trust in psychological studies but recognized more similarities between psychology and natural science fields. The study suggests that teaching about the replication crisis can help students understand the complexities of scientific research and the importance of reproducibility. It also highlights the need for open discussion about scientific practices and the self-correcting nature of science. In Sage Journals, this study has green-open access, open-access publishing, open access materials, however no pre-registration was completed.

Relation to study. Chopik et al. (2018) aimed to examine the effects of a one-hour lecture designed to communicate the replication crisis and recommendations to increase reproducibility in psychological research (https://osf.io/h3fvc). While the one-hour lecture has been reliably and empirically supported, feasibility and time are an issue in this case. To summarize, my study will focus on Honours students and First-Year Undergraduate Psychology

students watching a 10-minute video and answering a pre-post survey pertaining to their attitudes and knowledge on the crisis.

The video. I recorded a short 10-minute video and created PowerPoint slides pertaining to lecture material from Chopik et al. (2018) (e.g., QRPs, ManyLabs Project, etc.) and discussed current Open Science principles, as discussed above. ManyLabs was a national project where labs from all around the world were asked to try and reproduce the effects from a series of studies. Some of these were very famous and classic effects in psychology, others were less known, and yet others were more recent. These studies were chosen because they were easy to run. However, the goal of this study was not to thoroughly teach students about the crisis. With Honours psychology students, the goal was to examine their attitudes towards the replication crisis in psychological science after watching this short fundamental video and observe a change in their behavior with engaging in Open Science principles for their honours thesis at the end of the academic year. With First-Year Undergraduate Psychology students, the goal was to examine their attitudes towards the replication crisis in psychological science after watching this short fundamental video and observe a change in their perception to engage in proper Open Science practices.

The Crisis and How it is taught at the University of Manitoba

There are several prestigious quantitative professors at the University of Manitoba who collectively teach students about this replication crisis. Some professors teach it in *PSYC 2250-Introduction to Psychological Research*, *PSYC 2260-Introduction to Research Methods,* and others teach it in Honours Psychology courses such as *PSYC 3200-Thinking Critically About Psychological Research* and *PSYC 3340- Design and Analysis for Psychological Experiments*. While this may not seem like an alarming issue, from a research perspective, it can be

detrimental. Out of 7 professors who teach 2260, 2 of them teach the fundamentals of Open Science. Furthermore, even though the *PSYC 2250* textbook teaches replication crisis and open science principles in two chapters and in detail, there is no guarantee students will absorb and recollect the information come their Honours degree. There may be no direct effect, but in the long run, not performing the proper Open Science practices could be detrimental for their future in academia/research.

Present Study

The replication crisis in psychology has revealed that many published findings fail to hold up under independent scrutiny, highlighting a critical need for transparent and reproducible research practices. This study investigated how awareness of the replication crisis predicted Honours Psychology students' attitudes of the replication crisis and psychology, and how it was associated to their behavior to engage in Open Science practices for their honours thesis. Furthermore, I investigated First-Year Undergraduate Psychology students' attitudes of the replication crisis and psychology, and how it was associated to their perception to engage in proper Open Science practices in the future. I employed a pre-test post-test educational video intervention. Both groups completed a pre-video questionnaire assessing their attitudes toward psychology, the replication crisis, and intentions to use open science practices. They then watched a concise 10-minute video on the replication crisis and open science principles, followed immediately by a post-video questionnaire measuring shifts in those same attitudes. On a separate note, I asked students about the level of agreeableness with the crisis which inferred their willingness to engage in Open Science practices. In other words, does knowing about the crisis make someone more willing to engage in proper Open Science principles? The goal here was not to measure a pre-post difference of their knowledge on the crisis because (1) Honours

students have learnt the concept in previous years, and (2) First-Year Psychology students will learn about it in a second year psychology course (PSYC 2250); therefore, items relating to knowledge on the replication crisis was only available in the post-video questionnaire. Finally, I wanted to see if there was an interaction between the level of study (Honours vs. First-Year students) and attitudes towards psychology and the replication crisis on students' engagement in proper Open Science practices. At the end of the academic year, Honours students completed a follow-up survey reporting which open science practices they implemented in their thesis work.

Research Questions and Hypotheses

The proposed study examined these research questions (RQs) and hypotheses (H):

RQ1. Will the pre-post intervention (IV) predict Honours Psychology students' attitudes towards psychology and the replication crisis (DVs), which in turn will predict their engagement in proper Open Science practices at the end of the academic year?

H1. The pre-post intervention (IV) will predict Honours Psychology students' attitudes towards psychology and the replication crisis (DVs), which in turn will predict their engagement in proper Open Science practices at the end of the academic year.

RQ2. Will the pre-post intervention (IV) predict First-Year Psychology students' attitudes towards psychology and the replication crisis (DVs), which in turn will predict their perception of engagement in proper Open Science practices?

H2. The pre-post intervention (IV) will predict First-Year Psychology students' attitudes towards psychology and the replication crisis (DVs), which in turn will predict their perception of engagement in proper Open Science practices.

RQ3. How does knowledge on the replication crisis predict students' willing engagement in proper Open Science practices?

H3. Students who have a higher level of agreeableness that there is a replication issue will have more willing engagement in proper Open Science practices.

RQ4. How does engagement in proper Open Science practices differ between Honours Psychology and First-Year Undergraduate Psychology students from pre-video to post-video video?

H4. There will be a more significant increase in engagement in proper Open Science practices pre-video to post-video for Honours students than for First-Year students, reflecting a TimePoint x CourseLevel interaction.

Method

Participation Information and Exclusions

168 participants were recruited from PSYC 4520-Honours Research Seminar (n = 5) and PSYC 1200-Introduction to Psychology at the University of Manitoba (n = 163). Before starting the study, participants were asked to report basic demographic information such as age, ethnic or cultural origins, gender identity, year in school and university major (this question will target more towards the first years). Participant entries were excluded if they met any of the following exclusion criteria: 1) if students weren't registered in PSYC 4520 or PSYC 1200; 2) if they did not complete the demographic questions; 3) if they did not complete the pre-post surveys; 4) if they did not watch the 10-minute video; and 5) if they fail an attention check embedded in the survey measures in which they were asked to indicate a specific answer from a list of choices. PSYC 1200 students were compensated with course credit for participating through the SONA system. PSYC 4520 students' names were drawn to win one of three pre-paid visas. This is a good incentive without directly linking compensation to participation.

Research Design

As a research design, I conducted a pre-test post-test educational video intervention. Relying on previous literature to determine a potential effect size can be overestimated (Ioannidis, 2008; as seen in Whitt, 2022), therefore I aimed to achieve enough power to detect a smaller effect. To achieve 90% power for detecting a medium effect of d = 0.30 using a mixed-designs ANOVA, a total sample size of 50 participants was needed. I wanted to over-collect data in case of excluding multiple participants based on the exclusion criteria. Practically, using d = .30 protected against a Type II error if the true effect is smaller than earlier reports.

Analysis Plan

There were two main statistical analyses which tested Hypotheses 1, 2 and 4. First, a repeated-measures *t*-tests evaluated within-subjects change from pre-to post-video on Engagement in proper Open Science practices, Attitudes Toward Psychology, and Attitudes Toward the Replication Crisis to establish whether the intervention produced meaningful attitude and engagement shifts. Second, a 2 × 2 mixed-design ANOVA with level of study (Honours vs. First-Year) as the between-subjects factor and TimePoint (pre vs. post) as the within-subjects factor to examine group differences and the TimePoint × Level interaction, thereby determining whether Honours and First-Year students exhibited differential change across time. In addition, there were two separate analyses that tested Hypothesis 3. First, an independent samples *t* – test was conducted to compare Replication Crisis scores between PSYC 4520 and PSYC 1200 students. Second, a simple linear regression was conducted to examine whether Replication Crisis Knowledge significantly predicted Engagement in proper Open Science practices. Prior to analysis, it is important to note a very small sample size of Honours students (n = 5-6) and might not draw true meaningful conclusions about group differences, which can compromise the strength of hypothesis testing related to academic level.

Materials/Measures

This study had three main dependent variables: Engagement in proper Open Science Practices, Attitudes Towards Psychology and Attitudes towards the Replication Crisis. The independent variables were a Pre-Post Video Intervention and Replication Crisis Knowledge. These study measures (Attitudes Towards Psychology and Replication Crisis Knowledge) were adapted from those of Chopik et al. (2018), which are publicly available for use (https://osf.io/mh9pe/). I believe these study measures helped effectively answer the research questions. Each of the measures are described in more detail below. Measures were presented in

a specific order, and the items of each measure were randomized. The attention check questions (see Appendix A) were presented after the video to ensure students watched it, and demographic questions were presented before the first questionnaire.

Dependent Measures

Engagement in proper Open Science practices

Based on their attitudes, students answered survey questions developed by Beaudry et al. (2022). To measure norms and counternorms, Beaudry et al. (2022) focused on participants' evaluations of 10 specific open science practices that reflected diverse open science norms. The study encompassed questions concerning norms (how students felt research should be conducted), norms in practice (how students believe psychological research is conducted), and replicability (how replicable students believe psychological research is). This questionnaire helped infer on Honours students' behaviors to engage in Open Science practices (see Appendix B). To measure Honours students' behaviors in engaging in proper Open Science practices, I measured this by giving students an 8-item survey of what they engaged in for their thesis (i.e., pre-registration, power analysis, dissemination, etc.). This was completed at the end of the academic year. I also measured whether these were imposed by their advisor versus steps they took themselves (or perhaps some measure of agreement their advisor had with engaging in the practices?) (see Appendix C). There is a two-week period before students present their final thesis. I sent out a mass email to each Honours instructor asking them to make an announcement on UMLearn about the end of academic year survey. Instructors shared the survey link April 10th that directed them to the questionnaire in Qualtrics. Again, it was noted that surveys are anonymous, and that there are no correct answers. I merely wanted to measure their behavior by seeing if they engaged in proper Open Science practices.

Attitudes Toward Psychology

To measure students' attitudes toward psychology, students were asked to respond to five items that assessed their attitudes toward psychology (e.g., "I like psychology" and "I think psychology is a 'soft' science."; as seen in Whitt, 2022; see Appendix D). These questions were taken from Chopik et al. (2018). It is important to note that no psychometric data was reported, but the questionnaire was pre-registered.

Attitudes Toward the Replication Crisis

To measure students' attitudes toward the replication crisis, participants were asked to respond to six items that assessed their attitudes toward the replication crisis (e.g., "For a researcher, how important is choosing a sample size before running a study?" and "It is important for a researcher to disclose all measures and experimental conditions that were included in a study.", as seen in Whitt, 2022; see Appendix E). Although the items do not say 'Attitudes Toward the Replication Crisis', I assessed their attitudes based on how important they consider each item. The more important they consider each aspect of study designs, the more they know the crisis is a problem. These questions were taken from Chopik et al. (2018). It is important to note that no psychometric data was reported, but the questionnaire was pre-registered.

Independent Measures

Pre-Post Video Intervention

In this study, I employed a 10-minute pre-post intervention video aimed at enhancing students' attitudes towards psychology and the replication crisis, which in turn will help them engage in proper Open Science Practices.

Replication Crisis Knowledge

To measure students' knowledge of the replication crisis, participants were asked to respond with their level of agreement to seven items assessing issues surrounding replication (e.g., "Replication of research is only a problem in the field of psychology."; as seen in Whitt, 2022; see Appendix F). These questions were taken from Chopik et al. (2018). It is important to note that no psychometric data was reported, but the questionnaire was pre-registered.

Demographics

Demographic information such as age, ethnic or cultural origins, gender identity, year in school and university major were asked to student pre-video intervention.

Attention Check

To ensure students watched the video and could proceed with the second part of the study, students had to answer 2 specific questions pertaining to the video content. Students could not proceed with the study without selecting the correct answer.

Procedure

Following thesis proposal in November and ethics clearance in February, participant recruitment began in March. Undergraduate students enrolled in PSYC 4520 (Honours Research Seminar) and PSYC 1200 (Introduction to Psychology) were invited to participate in the study, which was conducted entirely online using Qualtrics. PSYC 4520 students were recruited through their course announcement platform on UMLearn, and PSYC 1200 students were recruited through an online study participation platform in psychology courses through the University of Manitoba called SONA. After providing informed consent, participants accessed a sequence of study materials beginning with demographic questions, followed by a pre-video

questionnaire. They then viewed a 10-minute educational video on the replication crisis, after which they completed attention check questions to ensure engagement; only those who answered correctly could proceed. The final stage involved a post-video questionnaire assessing changes in attitudes and behaviors in Engagement in proper Open Science practices. Data collection for the pre-test post-test video intervention concluded in late spring (May). Analyzing and writing results commenced shortly after.

End of Academic Year Survey

To assess how students' engagement in Open Science practices translated into actual behaviors during the thesis process, an eight-item self-report survey was developed and administered at the end of the academic year. The survey was designed to capture the extent to which Honours students engaged in key Open Science practices throughout their thesis project, as well as the role of their academic advisors in supporting or facilitating those practices. This measure asked students to retrospectively report whether they engaged in practices such as power analysis, preregistration, open data sharing, and plans for dissemination regardless of outcome. To ensure ecological validity, the questionnaire included conditional follow-up questions to gain additional insight into the platforms students used (e.g., OSF, AsPredicted, Zenodo) and the nature of advisor support. These items were designed to not only to index behaviors but also to explore the interpersonal and institutional influences that might scaffold – or hinder – proper Open Science implementation.

Pilot Data

To ensure the self-made educational video is valid, data was collected using a small, independent sample of students in PSYC 1200 (n = 5) and PSYC 3630 (n = 5) at the University

of Manitoba (N = 10). To ensure the video was reliable for this study, I ran Cronbach's alpha and McDonald's omega to examine whether the item scores (i.e., attitudes toward psychology and the replication crisis) are internally consistent within the same scale. Next, I ran an Exploratory Factor Analysis (EFA) to evaluate whether the items' scores are loaded onto one single factor (or multiple factors) and calculated the sum scores. Finally, I ran a repeated-measures t-test to examine whether there is a significant change on each of the sum scores (corresponding to each factor). If the p-values are significant, then this supports that the intervention significantly helped change/improve their engagement and/or attitudes. It is important to note that the pilot study's small sample size (N =10) limits generalizability, therefore results should be proceeded with caution.

Descriptive Statistics

Participants were collected from a small, independent sample of PSYC 1200-*Introduction to Psychology* (n = 5) aged between 18-25 years old and *PSYC 3630-Psychological Measurement and Assessment* aged between 18 to 45 years old (n = 5, M_{age} = 28.9). The ratio of men to female students in 3630 was 1:4, and ethnicity ranged from White/European, Southeast Asia, Indigenous and Black. One student was in their third year of studies, three in their fourth year of their studies and one in their fifth year of study. The ratio of men to female students in 1200 was 2:3, and ethnicity ranged from East Asia, Arab/West Asian and Black. All students were in their first year of studies.

Repeated Measures *t*-test

A repeated-measures *t*-test was conducted for all dependent variables to see if there is a significant difference in participants scores pre-to-post replication crisis video. Additionally, calculating the sum of scores for both groups of participants is important because it helps

quantify overall changes in the dependent variables before and after watching the video (see Figure 2). It is important to note that lower scores represent a reduced likelihood of agreement with items on the scale (i.e., 1 = *Strongly Disagree*).

PSYC 3630 Pre-Post Video Scores

A repeated-measures *t*-test was conducted to compare PSYC 3630 students' Attitudes Toward Psychology before and after watching the replication crisis video. Scores did not differ substantially between pre-video ($M = 23.17$, $SD = 10.91$) and the post-video ($M = 21.83$, $SD = 10.32$). The difference was not statistically significant, $t(4) = 1.43$, $p = .227$, although the effect size was large (Cohen's $d = .89$), reflecting considerable variability in this small sample. A repeated-measures *t*-test was conducted to compare PSYC 3630 students' Attitudes Toward the Replication Crisis before and after watching the replication crisis video. Scores did not differ substantially between pre-video ($M = 24.67$, $SD = 11.76$) and the post-video ($M = 23.83$, $SD = 11.27$). The difference was not statistically significant, $t(4) = .79$, $p = .473$, Cohen's $d = .426$. The effect size was small, indicating the difference was little to none. A repeated-measures *t*-test was conducted to compare PSYC 3630 students' Engagement in proper Open Science practices before and after watching the replication crisis video. Scores differed substantially between pre-video ($M = 54.67$, $SD = 27.98$) and the post-video ($M = 18.5$, $SD = 9.63$). The difference was statistically significant, $t(4) = 10.39$, $p < .001$, Cohen's $d = 3.76$. The effect size was very large, indicating that students' Engagement in proper Open Science practices decreased markedly following the video.

PSYC 1200 Pre-Post Video Scores

A repeated-measures *t*-test was conducted to compare PSYC 1200 students' Attitudes Toward Psychology before and after watching the replication crisis video. Scores did not differ

substantially between pre-video (M = 18.17, SD = 8.47) and the post-video (M = 19.68, SD = 9.78). The difference was not statistically significant, $t(4) = -1.23$, $p = .286$, Cohen's $d = -.386$. The effect size was small, indicating the difference was little to none. A repeated-measures t-test was conducted to compare PSYC 1200 students' Attitudes Toward the Replication Crisis before and after watching the replication crisis video. Scores did not differ substantially between pre-video (M = 20, SD = 10.30) and the post-video (M = 18.5, SD = 9.63). The difference was not statistically significant, $t(4) = 1.86$, $p = .137$, Cohen's $d = .293$. The effect size was small, indicating the difference was little to none. A repeated-measures t-test was conducted to compare PSYC 3630 students' Engagement in proper Open Science practices before and after watching the replication crisis video. Scores differed substantially between pre-video (M = 51.67, SD = 25.30) and the post-video (M = 23.83, SD = 11.27). The difference was statistically significant, $t(4) = 13.91$, $p < .001$, Cohen's $d = 6.32$. The effect size was very large, indicating that students' Engagement in proper Open Science practices decreased markedly following the video.

Results

Sample Characteristics

168 participants were recruited from PSYC 4520-*Honours Research Seminar* (n = 5) aged between 18-25 years old and PSYC 1200-*Introduction to Psychology* at the University of Manitoba (n = 163) aged between 18-55 years old (i.e., most students between the ages of 18-25 years old, M_{age} = 21.96. For PSYC 4520, four students who identified as female participated in the study, and one student who identified as non-binary/third gender participated in the study. Four students described their ethnic background as White/European (e.g., English, French, Scottish, Polish), and one as Indo-Caribbean. For PSYC 1200, 106 students who identified as female (60%) participated in the study, whereas only 55 students who identified as male (33%)

participated in the study, and two students who identified as non-binary/third gender (7%) participated in the study. Most ethnic backgrounds were White/European (e.g., English, French, Scottish, Polish), but there were a variety of backgrounds selected (see Appendix G). Participant entries were excluded if they met any of the following exclusion criteria: 1) if students were not registered in PSYC 4520 or PSYC 1200; 2) if they did not complete the demographic questions; 3) if they did not complete the pre-post surveys; 4) if they did not watch the 10-minute video; and 5) if they fail an attention check embedded in the survey measures in which they were asked to indicate a specific answer from a list of choices.

Reliability Analysis

Cronbach's Alpha values for variables of Attitude Toward Psychology and Attitudes Toward the Replication Crisis were between 0.5 – 0.6 which represented moderate reliability/internal consistency of these items. However, Engagement in proper Open Science Practices were between 0.6 – 0.9 in R, and Replication Crisis Knowledge were between 0.6-0.7, which represented moderate to high reliability/internal consistency of these items. While Cronbach's Alpha is widely used to assess internal consistency, it assumes equal factor loadings which can be assumed restrictive. McDonald's Omega is a more advanced reliability measure that does not assume equal item contributions. In other words, omega accounts for factor loadings which are more reliable here because the number of items differ for each variable (i.e., 6 items compared to 8 items). Omega values were coded through R. Omega hierarchical was interpreted to determine if the scale is unidimensional or if there are subfactors that contribute significantly (i.e., multidimensional). In general, omega hierarchical coefficients were relatively moderate (i.e., fell between 0.5-0.8). This indicates there is a general factor influence, but subfactors matter. I noticed that omega hierarchical values diminished post-video. This may

indicate that responses are less cohesive, meaning participants interpreted questions more differently post-intervention than they did pre-intervention or that items pre-video had a better response rate. This could mean students' attitudes and engagement improved/changed post-video. Performing an EFA helped me determine if items fall under one or multiple factors. For all reliability coefficients, see Table 1.

Factor Structure

After running an EFA with one and two fixed factors, I decided to retain one factor for Attitudes Toward Psychology pre-and-post video (see Table 2). Most of the proportion of variance was explained with one factor. The proportion of variance explained pre-video for Attitudes Towards Psychology was 76%, and 68% post-video. It is important to note some problematic items. Item 4 ("*I think psychology is a soft science"*) did not load onto a factor pre-video but loaded onto two factors post-video. Item 3 ("*I think psychological research is similar to research in fields like philosophy, literature, or modern languages"*) and Item 6 ("*I trust the results of studies done by psychologists"*) did not load onto a factor post-video. After running an EFA with one and two fixed factors, I decided to retain one factor for Attitudes Toward the Replication Crisis pre-and post-video (see Table 3). Most of the proportion of variance was explained with one factor. The proportion of variance explained pre-video for Attitudes Towards the Replication Crisis was 64% and 51% post-video. No problematic items were seen.

After running an EFA with one and two fixed factors, I decided to retain one factor for Engagement in proper Open Science Practices (see Table 4) pre-and-post video. However, it is important to note that some items were problematic pre-and-post video such as Item 2 (barely; Factor 1 = -.30), Item 6 (barely; Factor 1 = .31 in post-video), Item 7 (barely; Factor 1 = .30), Item 10, and Item 14. Most items loaded onto one factor with a cut-off exceeding .32 (Miller &

Lovler, 2018), but one item loaded onto two factors. Typically, factors with eigenvalues greater than 1 are retained (Larsen & Warne, 2010). In this case, none of the eigenvalues exceeded 1, which suggests one factor might be sufficient to explain most of the variance compared to the single variable. On the other hand, the proportion of variance explained by one factor pre-video for Engagement in Open Science practices was 59% and post-video was 55%. I decided to keep items to one factor because of the low sample size of Honours students but bolded the double-loaded items in the table. After running an EFA with one and two fixed factors, I decided to retain one factor for Replication Crisis Knowledge (see Table 5) post-video only. There were no problematic items, and 5 out of 7 items loaded into a single factor. Most of the proportion was explained with one factor (i.e., 66%).

Main Analyses

Pre-Post Intervention Predictions

Repeated-Measures *t*-test. A repeated-measured *t*-test examined changes in within-subject participants' Engagement in Proper Open Science Practices, Attitudes Toward Psychology, and Attitudes Toward the Replication Crisis pre- to -post intervention, assessing how questionnaire responses evolved over time (i.e., TimePoint). A repeated-measures *t*-test was conducted to compare the difference between Engagement in proper Open Science practices and Attitudes Toward Psychology at Time 1 (pre-test) and Time 2 (post-test). Results indicated a significant difference in mean scores, $t(163) = 75.43$, $p < .001$ from Time 1 ($M = 68.69$, $SD = 6.30$) to Time 2 ($M = 25.59$, $SD = 3.65$). The effect size, $d = -9.01$, indicated a very large effect size. These findings supported my first and second hypothesis that the intervention predicted students' Attitudes Toward Psychology pre-post intervention, which in turn predicted their Engagement in proper Open Science practices. A repeated-measures *t*-test was conducted to

compare the difference between Engagement in proper Open Science Practices and Attitudes Toward the Replication Crisis at Time 1 (pre-test) and Time 2 (post-test). Results indicated a non-significant difference in mean scores, $t(164) = -.844$, $p = .399$ from Time 1 ($M = 25.40$, $SD = 4.39$) to Time 2 ($M = 25.59$, $SD = 4.09$). The effect size, $d = -4.48$, indicated a very large effect size. These findings did not support my first and second hypothesis that the intervention predicted students' Attitudes Toward the Replication Crisis pre-post intervention, which in turn predicted their Engagement in proper Open Science practices.

Knowledge and Engagement Relationship

Descriptive Statistics. Descriptive analyses were conducted for the composite variable POSRCKQ_sum (i.e., post-video sum of Replication Crisis Knowledge items), which represented the sum scores of 165 participants on seven Replication Crisis Knowledge questions. Sum scores retain the full range of possible values, which can be useful for detecting variability, especially in longer scales. Results indicated that scores ranged from 20 to 40 ($M = 27.57$, $SD = 3.65$), and a median of 28, suggesting a relatively symmetrical distribution, which supports assumptions of normality, and a central tendency slightly skewed toward higher scores (i.e., higher agreeableness). The interquartile range (Q1 = 25, Q3 = 30) revealed that the central 50% of responses were closely clustered, indicating limited dispersion and consistency among participants.

Three missing values were identified and excluded using listwise deletion, a commonly accepted method in psychological and behavioral research when missingness is minimal and random (Tabachnik & Fidell, 2019). A visual inspection of the distribution (e.g., histogram) suggested no substantial skewness or extreme outliers, supporting the appropriateness of using mean and standard deviation to summarize the data (Field, 2018). This histogram showed a symmetrical distribution of the POSRCKQ_sum scores, with a clear peak around a score of 25 –

also the mode, based on frequency (see Figure 3). Most scores fall between 22 and 32, aligning nicely with the earlier descriptive statistics in the previous paragraph, reinforcing that the data cluster around the center with no dramatic skew or outliers (see Figure 3). This indicates that most participants tended to respond neutrally or slightly positively. These results suggest that the variable demonstrates sufficient normality and internal coherence to proceed with an independent samples-*t*-test, and linear regression model.

Independent-Samples t-test. An independent samples *t*-test was conducted to compare POSRCKQ_sum scores between students in PSYC 4520 and PSYC 1200. The results indicated no significant difference in scores between the two courses, $t(163) = -.73$, $p = .468$. The mean score for PSYC 4520 ($M = 26.40$, $SD = 1.95$) was slightly lower than that of PSYC 1200 ($M = 27.61$, $SD = 3.68$), but this difference was not statistically meaningful. There wasn't a large difference in higher agreeableness between courses. The 95% confidence interval for the mean difference ranged from -4.48 to 2.07, indicating that the true difference could plausibly favor either group. However, the effect size, as measured by Cohen' $d = -.33$, suggests a small to moderate practical difference in scores (Cohen, 1988). This implied that while the statistical test didn't detect significance, the group difference may still carry interpretive value in the context of replication knowledge and course experience.

Linear Regression Model. A simple linear regression was conducted to examine whether Replication Crisis Knowledge (POSRCKQ_sum) significantly predicted Engagement in proper Open Science practices (DEBQ_sum). The model was statistically significant, $F(1, 161) = 11.55$, $p = .001$, with an R^2 of .07, indicating that approximately 7% of the variance in Open Science Engagement was explained by replication knowledge scores. The unstandardized regression coefficient was $\beta = .86$, $SE = .25$, $t = 3.40$, $p = .001$. This suggests that for every one-unit increase

in Replication Knowledge, Engagement in proper Open Science practices increased by .86 points, on average. The strength and direction of the relationship indicate a positive association between students' understanding of the replication crisis and their willing engagement in Open Science practices. While the effect size was small-to-moderate (d = .27), the relationship was statistically reliable and consistent with theoretical expectations about knowledge influencing behavior. This result supported the hypothesis that greater knowledge of the replication crisis is positively associated with students' willing Engagement in proper Open Science Practices. However, the effect size was small-to-moderate, and because the analysis was correlational, causal inferences cannot be drawn.

Course Level Interaction

A 2 (Course Level: First-Year vs. Honours) x 2 (TimePoint: Pre vs. Post) mixed-design ANOVA was conducted to examine whether Engagement in proper Open Science practices varied as a function of students' level of study over time. Students were removed if questions were not answered, which resulted in 5 Honours Students and 158 First Years. The within-subjects factor was TimePoint (pre-video vs. post-video), and the between-subjects factor was Course Level. There was no significant main effect of TimePoint, $F(1, 320) = .348$, $p = .56$, $\eta^2_p = .001$, suggesting that engagement scores did not significantly change from pre- to post- intervention when collapsed across course groups, and the effect size was negligible. However, there was a significant main effect of Course Level, $F(1, 320) = 4.658$, $p = .032$, $\eta^2_p = .014$, suggesting an overall difference in engagement between Honours and First-Year students, with a small effect size. Finally, the TimePoint x Course Level interaction was not significant, $F(1, 320) = .004$, $p = .96$, $\eta^2_p = .000$, indicating that the pattern of change in engagement over time was not different between the two course levels, and the effect size was negligible. This result failed to support Hypothesis 4, which

predicted a significant interaction between course level and time. Across analyses, observed effect sizes were negligible to small (η^2p range = .000–.014), indicating that differences in engagement were statistically detectable but limited in magnitude.

Follow-up Analyses

End of Academic Year Survey

Of the 13 respondents, seven (53.8%) reported completing a power analysis, and six (46.2%) pre-registered their study – most using OSF or AsPredicted. One student indicated that preregistration was managed by their advisor, suggesting variability in student-led versus advisor-led implementation. Similarly, six students indicated they were planning to share their data openly, through several opted not to disclose this information. Five students (38.5%) reported having a publication or dissemination plan independent of study outcomes. Advisor involvement appeared to play a facilitative role: seven students noted that their advisor encouraged Open Science practices, and five indicated their advisor provided hands-on guidance throughout the process. These findings suggest that while formal engagement in Open Science may have been influenced by earlier components of the study (e.g., the intervention and replication knowledge), actual behavioral engagement during thesis completion was meaningfully shaped by advisor support and applied practice. The presence of key Open Science behaviors – especially among Honours undergraduates – reflects emerging patterns of adoption and reinforces the idea that reform-minded training can translate into concrete scholarly behaviors when scaffolded by mentorship.

Discussion

Summary

This study aimed to evaluate an educational intervention on students' attitudes, knowledge, and self-reported behaviors related to the replication crisis and Open Science practices. Undergraduate psychology students from both Honours and First-Year courses participated in a pre-test/post-test design, with Honours students also completing a follow-up survey about the Open Science methods they used in their thesis projects. The intervention led to meaningful changes in students' Attitudes toward Psychology and their Engagement in proper Open Science practices. However, no significant change was found in Attitudes toward the Replication Crisis. Replication Crisis Knowledge was similar across both academic levels, and students who demonstrated greater understanding of replication issues tended to report more willing Engagement in proper Open Science practices. The video appeared to be consistent across both Honours and First-Year students, suggesting that academic level alone was not associated to how students responded to the pre-test post-test intervention. Among Honours students, many reported using Open Science strategies such as power analysis, pre-registration, and open data sharing in their thesis work. Advisor support seemed to play a meaningful role in encouraging these behaviors. Together, these findings highlighted the importance of not only educating students about Open Science but also providing mentorship and practical opportunities to apply these principles in their own research.

Interpretations

Several findings across pre-post comparisons, groups differences, and predictive models helped clarify the scope and limitations of the intervention's effectiveness. A significant decrease in self-reported Attitudes toward Psychology and Engagement in proper Open Science practices

was observed following the intervention. Although initially counterintuitive, this may reflect a recalibration of students' self-perceptions after gaining a clearer understanding of what constitutes "proper" Open Science behavior. As Chopik et al. (2018) noted, increased exposure to replication-related issues can heighten students' awareness of methodological shortcomings, which may lead to greater critical self-reflection and, in turn, lower self-ratings of compliance. In this sense, the intervention may have introduced a more rigorous internal benchmark for what counts as authentic engagement with reform practices. In contrast, students' self reported Attitudes toward the Replication Crisis and Engagement in proper Open Science practices remained largely unchanged. This attitudinal stability may suggest that such beliefs/attitudes about the crisis are relatively stable and may not shift meaningfully following a single, brief intervention. Chopik et al. (2018) similarly found only modest movement in attitudes after a one-hour lecture and suggested that more immersive instructional formats – such as semester-long discussions or reflective assignments – may be necessary to elicit deeper attitudinal change.

Descriptive analyses of Replication Crisis Knowledge indicated that students had a reasonably well-distributed understanding of the topic, with minimal differences between First Year and Honours students. Although group differences were not statistically significant, a small-to-moderate effect size suggests there may be some practical divergence based on academic exposure. According to Chopik et al. (2018), replication topics are increasingly included in lower-level classes, especially when instructors purposefully incorporate reform-oriented issues into the curriculum. At the University of Manitoba, students learn about replicability and its issues in their second-year introduction to research methods course by conducting their own hypothetical research study. This activity provides them with a practical way to see how research findings can vary and why repeating studies is important. It also encourages careful thinking

about how studies are designed and communicated. By introducing these ideas early, the program helps students develop an appreciation for the value of transparency and openness in psychological science, which they can carry forward into more advanced coursework for their future research. Importantly, greater Replication Crisis Knowledge significantly predicted increased self-reported Engagement in proper Open Science practice behaviors. This relationship supports theoretical assumptions that knowledge can foster behavioral intention, a position supported by Chopik et al. (2018), who advocated for embedding replication-related content into undergraduate training to foster methodological awareness. While the effect size was modest, the relationship was statistically reliable, suggesting that conceptual understanding may play an influential – though not exclusive – role in shaping students' scientific behavior.

Finally, the hypothesized interaction between course level and timepoint was not supported. The educational intervention didn't differently predict Engagement in proper Open Science practices for First-Years versus Honours students. Several explanations may account for this finding. First, ceiling effects among Honours students may have limited the potential for further change. Second, the intervention itself may not have been sufficiently leading or sustained to produce measurable effects, particularly across heterogeneous groups (Chopik et al., 2018). Third, the engagement scale may not have been sensitive enough to detect subtle behavioral shifts. Fourth, First Year students may have already been exposed to replication crisis discourse through their introduction to psychology course or other channels, narrowing the expected gap between groups. Finally, statistical limitations such as reduced variability and contextual factors like academic stress or cognitive load may have been confounding variables which could have predicted the outcomes. Each of these interpretations aligns with Chopik et al.

(2018) findings that educational impact depends not only on content but also on delivery, timing, and students' preexisting beliefs.

Crucially, the end-of-year behavioral survey provided complementary evidence that many students did, in fact, implement core Open Science practices in their thesis work. Over half reported completing a power analysis and preregistering their study, and nearly half indicated plans to share data and disseminate their results regardless of outcome. Furthermore, more than half reported that their advisor encouraged Open Science, and a substantial portion received direct guidance through these processes. These findings suggest that for at least some students, conceptual understanding translated into applied behavior – particularly when scaffolded by advisor support. This behavioral uptake helps validate the earlier regression findings and demonstrates that even modest instructional interventions, paired with mentorship, can foster meaningful engagement with scientific reform practices at the undergraduate level.

Implications

The findings from this study contribute to our understanding of how undergraduate psychology students engage with Open Science principles when provided with structured educational exposure and mentorship. Although the intervention did not significantly shift attitudes toward the replication crisis or reveal differences across academic levels, other outcomes point to promising avenues for encouraging a more reform-oriented research culture. The positive link between Replication Crisis Knowledge and Engagement in proper Open Science practices underscores the value of integrating replication education into undergraduate coursework. Early exposure may encourage lasting change, supporting calls to treat Open Science as a core competency rather than an optional add-on (Chopik et al., 2018). Evidence of behavioral follow-through – such as students conducting power analyses, preregistering studies,

and planning to share data – demonstrates that Open Science can be practices even at the undergraduate level, especially when guided by faculty mentors. These results highlight mentorship as a vital bridge between values and action, suggesting that training should extend to both students and supervisors for future research. For instructors, this means moving beyond simply teaching about replication to actively co-constructing Open Science practices with students.

Strengths

This study demonstrates several important strengths that enhance the credibility, relevance, and practical impact of its findings, ranging from the design of the intervention to the inclusion of diverse student groups and the connection between attitudes and real behaviors.

Pre-Post Design

One of the main strengths of this study is its use of a pre-post design, which allowed for the direct observation of changes in students' attitudes and behaviors. By measuring the same participants before and after the video intervention, the study was able to capture within-person differences rather than relying on comparisons between separate groups. This approach reduces the influence of individual differences, since each student serves as their own baseline. It also makes the findings more meaningful, because the changes observed can be tied to the intervention rather than to differences in who happened to be in each group. While cross-sectional designs can show differences between groups, they cannot demonstrate how individuals change, which is why the repeated-measures approach is so valuable here. The design also highlights the dynamic nature of attitudes, showing that they are not fixed and can shift even after a brief educational experience. This is especially important in psychology education, where interventions aim to shape both knowledge and practice. By using this design, the study provides

stronger evidence that the video changed attitudes, even if they were minimal. This sets the stage for suture work to build on these findings with longer interventions or follow-ups. Overall, the pre-post design is a clear strength because it captures meaningful change in a way that is both practical and informative.

Honours vs. First-Year Psychology Students

Another strength of the study is the inclusion of both Honours students and First-Year students, which allowed for comparisons across different levels of training. Honours students bring more advanced experience with research methods, while First-Year students are just beginning their psychology education. By examining both groups, the study was able to show how responses to replication education may differ depending on prior exposure and academic maturity. This comparison adds depth to the findings, since it highlights that attitudes toward replication and Open Science are not uniform across students. It also suggests that interventions may need to be tailored to different levels of study, with more advanced students perhaps ready to engage in practical applications while newer students are still forming their basic understanding. This choice of participants is also a foundation for future research to explore how attitudes evolve as students move through their degree programs. In short, the inclusion of both Honours and First-Year students adds richness to the study by allowing me to view the attitudes from one level of study to another.

Multiple Measures

The study's use of multiple measures is another important strength. Rather than focusing on a single outcome, the research examined Attitudes Toward Psychology, Attitudes Toward the Replication Crisis, Replication Crisis Knowledge, and Engagement in proper Open Science practices. This multidimensional approach provides a fuller picture of how students respond to

replication education. Attitudes alone can show how students feel, they do not necessarily reveal whether those feelings translate into action. By including measures of actual engagement, the study was able to connect attitudes to behaviors, which is crucial for understanding the impact of educational interventions. Looking at multiple outcomes also helps identify areas where change is more or less likely to occur. For example, students may shift their attitudes toward psychology but remain unchanged in their attitudes toward the replication crisis, or they may show strong changes in engagement behaviors. This level of detail makes the findings more nuanced and useful for educators who want to know which aspects of replication education are most effective. It also strengthens the validity of the study, since the results are not dependent on a single measure. Overall, the inclusion of multiple dependent variables makes the study more comprehensive and provides a richer understanding of how students respond to replication crisis education.

Accessible Intervention

The intervention itself is also a strength because it was brief, standardized, and easy to deliver. Using a short video allowed the study to reach many students in an efficient way. The standardized format ensures that all students received the same information, which increases consistency across participants. The accessibility of the video also makes it scalable, meaning it could be used across different courses or institutions. While longer or more interactive interventions may be valuable in the future, the success of this brief video demonstrates that small steps can still make a difference. It also highlights the potential of digital tools in psychology education, especially in online or blended learning environment. Overall, the accessible nature of the intervention is a strength because it shows that replication education can be delivered effectively in a way that is both practical and impactful.

Linking Attitudes to Behavior

Finally, a major strength of the study is that it went beyond measuring attitudes to examine students' actual behaviors of Open Science practices. Many studies stop at asking participants how they feel, but this research looked at whether students' engagement in Open Science practices such as pre-registration, power analysis, or data sharing by administering an 8-item self-report survey to Honours students at the end of the academic year. This connection between attitudes and behaviors is important because it shows whether changes in thinking translate into action. By linking attitudes to behavior, the study provides stronger evidence that replication education can predict students' future research practices. This is especially valuable for Honours students, who are already conducting independent projects and making decisions about how to design and report their work.

Limitations

While this study offers valuable insight into students' engagement with Open Science practices, some limitations should be acknowledged when interpreting the findings. First, the sample included a relatively small number of Honours students ($n = 6$) in the actual study, which may have limited statistical power to detect meaningful interaction effects between academic level and time. This imbalance could also obscure group difference specific trends, making it difficult to draw firm conclusions about how seniority or research experience may shape responsiveness to replication-related interventions. Additionally, the sample was composed exclusively of psychology undergraduates from a single academic institution, which comprised mostly of females. As such, the findings may not generalize to students in other disciplines or settings, particularly those with differing methodological norms, training models or exposure to Open Science values. This reduces the study's external validity, and random sampling should be

used in future studies. Psychology, as a field, has been at the center of the replication crisis discourse; thus, students may be more primed to engage with reform-oriented content than their peers in less affected domains.

Second, the repeated-measures *t*-test examining the difference in mean scores between Attitudes Toward Psychology and Replication Crisis, and Engagement in proper Open Science practices had a significantly higher effect size than usual. While this suggests a dramatic difference between the two variables, such a magnitude is atypical in psychological research and likely reflects differences in scale ranges or measurement properties rather than a true effect. The questionnaire assessing Attitudes Toward Psychology from Chopik et al. (2018) consisted of 6 questions on a 7-point Likert scale, the Attitudes Toward the Replication Crisis questionnaire also from Chopik et al. (2018) consisted of 6 questions on a 5-point Likert scale, and the Engagement in proper Open Science practices questionnaire from Beaudry et al. (2022) consisted of 20 questions on a 5-point Likert scale. Future work should ensure comparability measures (i.e., creating a new measurable scale with similar properties) and consider alternative effect size metrics that account for scale differences. As a result, this estimate from this study should be interpreted with caution and treated as exploratory. A further limitation concerns the measurement properties of several scales. Internal consistency was modest ($\alpha = 0.5$-0.6) for Attitudes Toward Psychology and Attitudes Toward the Replication Crisis, suggesting that items may not have reliably captured the intended constructs. Chopik et al. (2018) did not report psychometric data, but questions were limited (i.e., 6), therefore I did not want to remove any at this time in case there was no significant effect. Future studies need to focus on adding extra questionnaire items or removing existing ones if moderate reliability persists.

Finally, although the pretest-posttest comparisons provided evidence consistent with H1 and H2, the single-group design introduces threats to internal validity that limit the strength of these conclusions. For example, participants may have improved simply because of repeated exposure to the measures (testing threat), rather than due to the video intervention. Natural changes over time (maturation) or external experiences between the pretest and posttest (history) could also account for observed differences (Morling, 2021). This study conducted an online survey and could not control for extraneous variables. In addition, regression to the mean may explain changes among participants who scored unusually high or low at baseline. Because no control group was included, it is not possible to rule out these alternative explanations (Morling, 2021). Finally, the intervention itself was brief and standardized, which helps with internal consistency but may not reflect how replication education is delivered in diverse instructional contexts. Future studies might explore how pedagogical delivery methods, and advisor attitudes interact to shape the uptake of Open Science practices over time.

Future Directions

The findings of the present study open several avenues for future research aimed at advancing educational strategies to promote Open Science engagement among undergraduates. While the current work focused on psychology students within a single institution using a primarily survey-based design, further inquiry can refine, expand, and contextualize these results in meaningful ways such as sampling a more representative group of participants and performing a longitudinal study. Because this study focused exclusively on psychology students—who may already be primed to engage with replication-related discourse—future studies should explore students in other social science disciplines, including sociology, political science, education, and economics. These fields vary in methodological tradition, openness norms, and exposure to

replication debates. A comparative study could examine whether similar interventions yield divergent effects based on epistemological cultures, research traditions, or advisor norms. For example, preregistration may be seen as essential in economics but unfamiliar in anthropology. Understanding such disciplinary characteristics could help tailor future training materials for maximum understanding. Also, given the early-stage nature of reform adoption among undergraduates, longitudinal research could assess the *durability* of replication education. Tracking students across the course of their undergraduate degrees—or into graduate school—could clarify whether knowledge gained in a brief intervention leads to sustained behavioral adoption. Are students who engaged in Open Science as undergraduates more likely to carry those practices into future projects? Do early experiences with preregistration increase comfort or fluency with transparent methodologies over time? Longitudinal designs could also examine the evolving influence of advisor mentorship across stages of academic development.

Measurement issues also warrant attention. The unusually large effect sizes observed between Attitudes Toward Psychology, Attitudes Toward Replication Crisis, and Engagement in Open Science practices likely reflect differences in scale ranges and measurement properties rather than true effects. Future research should prioritize the development of comparable scales with consistent response formats and consider alternative effect size metrics that account for scale differences. Internal consistency was modest for the Chopik et al. (2018) scales, suggesting that items may not reliably capture the intended constructs. Refinement through items expansion or removal will be necessary to improve reliability.

Looking ahead, future research could strengthen internal validity by using designs that help separate the impact of the video from other factors that might influence students' responses. Even though the targeted groups were first years and Honours students, adding a control group or

randomizing participants would make it easier to tell whether changes are truly due to the intervention rather than repeated testing, natural shifts over time, or outside experiences. Following students over a longer period could also show whether these changes last or fade, and whether regression to the mean plays a role. It will also be important to consider and control for third variables – such as prior exposure to Open Science, general attitudes toward research, or differences in course experiences – that may shape how students respond to replication crisis education. Beyond design improvements, future studies should explore how replication education unfolds in real classrooms, through lectures, discussions, or mentorship, and how instructor's own attitudes toward Open Science influence students' engagement. In short, moving beyond brief, standardized interventions toward richer, more varied approaches will provide stronger evidence for how replication crisis education can emit a lasting change in student practices.

Conclusion

In sum, this study highlights the complex interplay between education, mentorship, and student behavior in promoting Open Science engagement among undergraduates. While brief educational interventions may not dramatically shift attitudes, they can recalibrate self-awareness and, when combined with mentorship, translate into meaningful behavioral change. Behavioral follow-through—such as preregistration, power analyses, and data sharing—was evident among a substantial portion of students, particularly when supported by advisor encouragement and guidance. These findings underscore the idea that fostering a culture of transparency in science requires not just content delivery but also ongoing mentorship and institutional support. The results offer an encouraging glimpse into the potential of undergraduate education to cultivate

scientifically responsible and reform-minded researchers, while also pointing to the need for sustained, context-sensitive strategies to support this transition.

References

Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. Emotion, *12*, 154 –162. http://dx.doi.org/ 10.1037/a0023527

Baur, C. (2022, January 26). *Explainer: What is Selective Reporting Bias?* Www.researchsquare.com. https://www.researchsquare.com/blog/what-is-selective-reporting-bias

Beaudry, J. L., Williams, M. N., Philipp, M. C., & Kothe, E. J. (2022). What do Incoming University Students Believe About Open Science Practices in Psychology? *Teaching of Psychology*, 9862832211002-. https://doi.org/10.1177/00986283221100276

Button K. S., Ioannidis J. P. A., Mokrysz C., Nosek B. A., Flint J., Robinson E. S. J., Munafò M. R. (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 1–12. Crossref. PubMed. ISI.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward republicanism up to 8 months later. Psychological Science, *22*, 1011–1018.

Chambers, C. D. (2017). The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. Princeton, NJ: Princeton University Press.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Erlbaum

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. Pharmaceutical Statistics, *10*, 311–317. http://dx.doi.org/ 10.1002/pst.467

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. eLife, *10*, e61701. https://doi.org/ 10.7554/eLife.71601

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Forbes, H. J., Travers, J. C., & Johnson, J. V. (2023). Supporting the replication of your research. *In Research Ethics in Behavior Analysis* (pp. 237–262). https://doi.org/10.1016/B978-0-323-90969-3.00003-7

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the Differences Between Bayesian and Frequentist Statistics.

International Journal of Radiation Oncology, Biology, Physics, *112*(5), 1076–1082. https://doi.org/10.1016/j.ijrobp.2021.12.011

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? Psychophysiology, *33*, 175–183. http://dx.doi.org/10.1111/ j.1469-8986.1996.tb02121.x

Hartgerink C H J, Wicherts, J. M., & van Assen M A L M. (2017). Too Good to be False: Nonsignificant Results Revisited. *Collabra. Psychology*, *3*(1). https://doi.org/10.1525/collabra.71

Ioannidis, J. (2005) Why most published research findings are false. *PLoS Medicine* 2: 696–701. Crossref. ISI.

Ioannidis, J. (2008). Why most discovered true associations are inflated. Epidemiology, 19(5), 640-648.

Jhangiani, R. S., & Chiang, I.-C. A. (2015). *Research Methods in Psychology - 2nd Canadian Edition* (2nd Canadian Edition). BCcampus.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnı´k, Sˇ., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability. Social Psychology, *45*, 142–152.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, *31*, 107–112. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00578.x

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior research methods*, *42*, 871-876.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). Social Psychology, *45*, 216–222.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological Methods, *9*, 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is Psychology Suffering From a Replication Crisis?: What Does "Failure to Replicate" Really Mean? *The American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Metskas, A. (2022, December 6). *How to fix the replication crisis in psychology*. Clearer Thinking. https://www.clearerthinking.org/post/our-new-project-to-help-fix-the-replication-crisis-in-psychology

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach* (6th ed.). SAGE Publications.

Morling, B. (2021). *Research methods in psychology : evaluating a world of information* (Fourth edition.). W.W. Norton & Company.

Nikolopoulou, K. (2023, February 15). *What Is Publication Bias? | Definition & Examples.* Scribbr. Retrieved June 3, 2024, from https://www.scribbr.com/research-bias/publication-bias/

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

Pashler H., Harris C. R. (2012) Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science* 7: 531–536. Crossref. PubMed. ISI.

Renkewitz, F., & Heene, M. (2019). The Replication Crisis and Open Science in Psychology: Methodological Challenges and Developments. *Zeitschrift Für Psychologie*, *227*(4), 233–236. https://doi.org/10.1027/2151-2604/a000389

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin, *113*, 553–565. http://dx.doi.org/10.1037/0033-2909 .113.3.553

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

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, *47*, 1173–1181. http://dx.doi.org/10.1037/0003-066X.47 .10.1173

Schmidt S. (2009) Shall I really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* 13: 90–100. Crossref. ISI.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. Psychological Methods, *3*, 403– 411. http://dx.doi.org/10.1037/1082-989X.3.4.403

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, *40*, 73– 83. http://dx.doi.org/10.1037/0003-066X.40.1.73

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology. Psychological Science, 22, 1359–1366. https://doi.org/10.1177/0956797611417632

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. American Statistician, 49, 108–112. https://doi.org/10.2307/2684823

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

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson Education.

Whitt, C. M. (2022). *THE EFFECTS OF TEACHING ABOUT THE REPLICATION CRISIS ON UNDERGRADUATES' EPISTEMIC DEPENDENCE* (pp. 1–120) [DISSTERTATION *THE EFFECTS OF TEACHING ABOUT THE REPLICATION CRISIS ON UNDERGRADUATES' EPISTEMIC DEPENDENCE*]. The Effects of Teaching About The Replication Crisis on Undergraduates Epistemic Dependence.pdf

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. Science, *322*, 606–607.

Wortman, J., Donnellan, M. B., & Lucas, R. E. (2014). Can physical warmth (or coldness) predict trait loneliness? A replication of Bargh and Shalev (2012). Archives of Scientific Psychology, *2*, 13–19. http:// dx.doi.org/10.1037/arc0000007

» About. (2024). Clearerthinking.org. https://replications.clearerthinking.org/about/

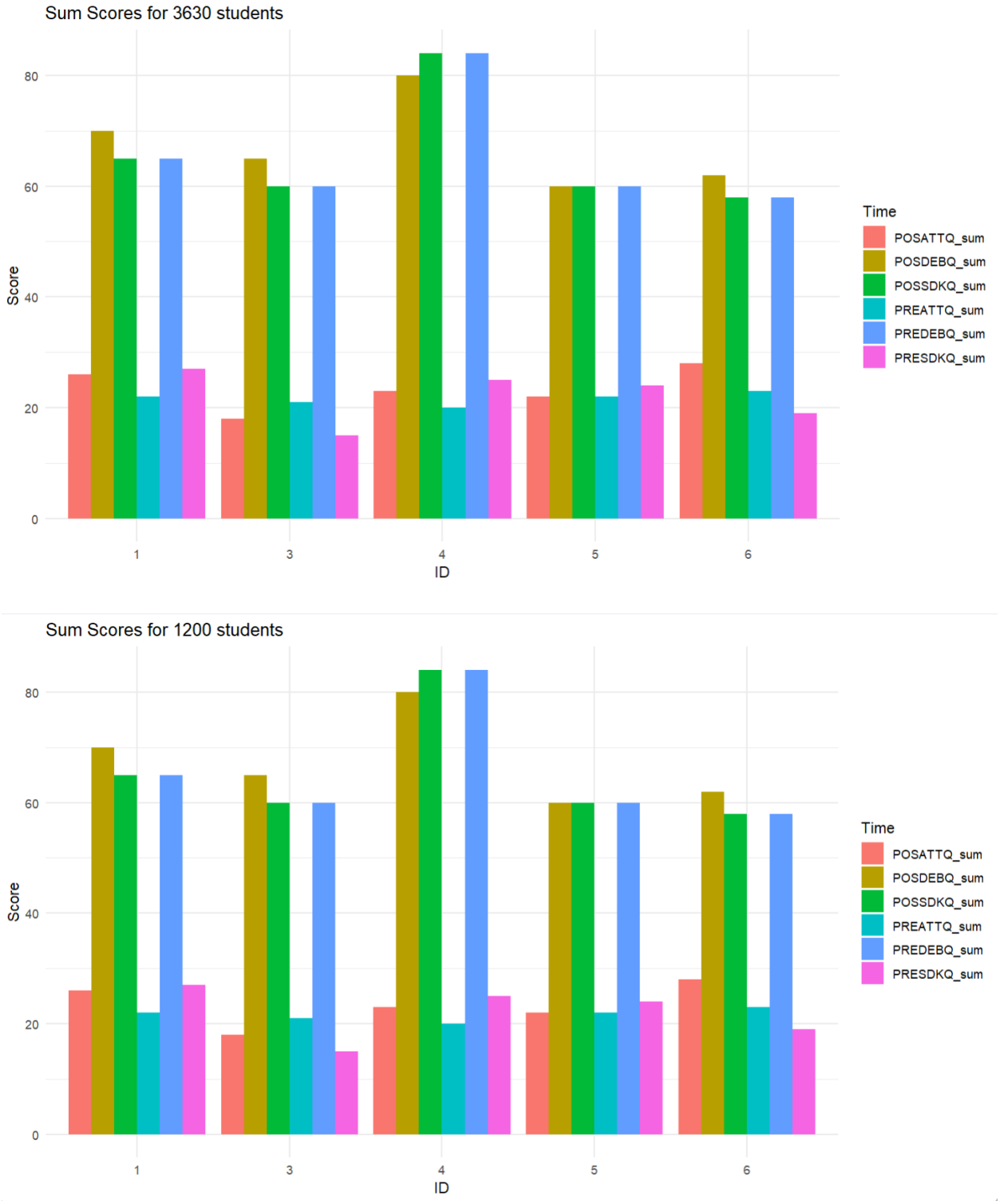
Open Science Practices and Methods: What They Are, Why They Are Important, and Examples. (2023, March 3). Orvium. https://blog.orvium.io/open-science-practices/

Science, C. for O. (n.d.). *Preregistration*. Www.cos.io. https://www.cos.io/initiatives/prereg

Figure 1

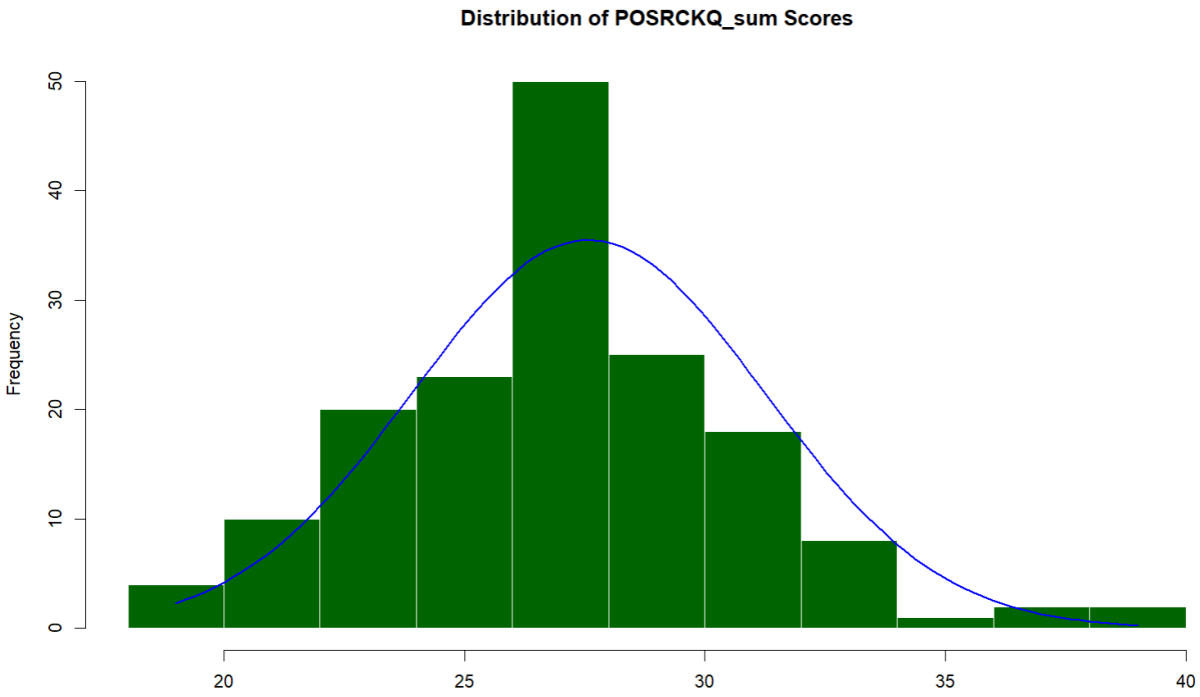
Note. Digital Badges from the Center for Open Science (as seen in Jhangiani & Chiang, 2015)

Figure 2


Sum Scores for 3630 students
Score
ID
Time
POSATTQ_sum
POSDEBQ_sum
POSSDKQ_sum
PREATTQ_sum
PREDEBQ_sum
PRESDKQ_sum
Sum Scores for 1200 students
Score
ID
Time
POSATTQ_sum
POSDEBQ_sum
POSSDKQ_sum
PREATTQ_sum
PREDEBQ_sum
PRESDKQ_sum

Note. $N = 10$. This figure represents both sample (Honours and First-Years) of participants' scores pre-post video for different variables. PRE refers to pre-video items and POS post-video items. ATTQ refers to attitudes (i.e., Attitudes Toward Psychology). SDK refers to study design knowledge (i.e., Attitudes Toward the Replication Crisis). DEB refers to the name Debra which was used hypothetically in the Norm and Counternorms Items (i.e., Engagement in proper Open Science practices).

Figure 3



Note. Histogram displaying the sum of scores for the variable Replication Crisis Knowledge Post-Intervention.

Table 1

Cronbach's Alpha and McDonald's Omega Reliability Coefficients

Items	Cronbach's α	McDonald's Ω Hierarch.
PREATTQ (1 – 6)	.56	.56
POSATTQ (1 – 6)	.52	.36*
PRESDKQ (1 – 6)	.62	.59
POSSDKQ (1 – 6)	.65	.47
PREDEBQ (1 – 20)	.66	.67
POSDEBQ (1 – 20)	.82	.49
POSRCKQ (1-7)	.71	.69

Note. Asterisk* represents relatively low reliability coefficients. PRE refers to pre-video items and POS post-video items. ATT refers to attitudes (i.e., Attitudes Toward Psychology). SDK refers to study design knowledge (i.e., Attitudes Toward the Replication Crisis). DEB refers to the name Debra which was used hypothetically in the Norm and Counternorms Items. RCK refers to Replication Crisis Knowledge. All Q's represent the word 'question'. 'Hierarch.' represents Omega Hierarchical.

Table 2

EFA for Attitudes Toward Psychology

Items	Factor 1	Factor 2
PREATTQ1	.89	-.05
PREATTQ2	.43	.19
PREATTQ3	-.01	-.54
PREATTQ4	-.25	.17
PREATTQ5	.51	.24
PREATTQ6	.33	-.10
POSATTQ1	.78	-.10
POSATTQ2	.24	-.42
POSATTQ3	.30	.28
POSATTQ4	-.01	.53
POSATTQ5	.63	.15
POSATTQ6	.26	-.03

Note. Exploratory Factor Analysis (EFA) extraction using Principal Axis Factoring (PAF) as the factor extraction method and oblimin rotation. PREATTQ4, POSATTQ3, and POSATTQ6 didn't load into any factor.

Table 3

EFA for Attitudes Toward the Replication Crisis

Items	Factor 1	Factor 2
PRESDKQ1	.57	-.08
PRESDKQ2	.47	-.07
PRESDKQ3	.47	.04
PRESDKQ4	.67	.05
PRESDKQ5	.00	.82
PRESDKQ6	.35	.20
POSSDKQ1	.43	.24
POSSDKQ2	-.04	.75
POSSDKQ3	.13	.48
POSSDKQ4	.84	.00
POSSDKQ5	.09	.44
POSSDKQ6	.42	-.15

Note. Exploratory Factor Analysis (EFA) extraction using Principal Axis Factoring (PAF) as the factor extraction method and oblimin rotation. Bolded values represent double-loaded items.

Table 4

EFA for Engagement in Proper Open Science Practices

Items	Factor 1	Factor 2
PREDEBQ1	.40	-.01
PREDEBQ2*	-.27	.12
PREDEBQ3*	-.22	-.15
PREDEBQ4	.42	.15
PREDEBQ5*	.14	.16
PREDEBQ6*	.12	-.04
PREDEBQ7*	.03	.17
PREDEBQ8*	-.21	.02
PREDEBQ9	.75	.10
PREDEBQ10*	-.16	-.01
PREDEBQ11	-.41	-.16
PREDEBQ12	.62	.03
PREDEBQ13	.61	-.22
PREDEBQ14*	-.25	.24
PREDEBQ15	.00	-.61
PREDEBQ16	.20	.59
PREDEBQ17	.34	-.47
PREDEBQ18	-.04	.54
PREDEBQ19	-.11	-.39
PREDEBQ20	.35	.02

POSDEBQ1	**.**68	.09
POSDEBQ2*	**-**.30	.10
POSDEBQ3	-.18	.47
POSDEBQ4	.58	-.14
POSDEBQ5	-.04	.49
POSDEBQ6*	.31	-.18
POSDEBQ7*	.30	-.05
POSDEBQ8	-.26	.41
POSDEBQ9	.70	.01
POSDEBQ10*	**-**.24	.25
POSDEBQ11	-.34	.40
POSDEBQ12	.67	-.04
POSDEBQ13	.42	.07
POSDEBQ14*	.07	.11
POSDEBQ15	**-**.11	.71
POSDEBQ16	**.36**	**-.44**
POSDEBQ17	.30	**-**.71
POSDEBQ18	.12	-.38
POSDEBQ19	-.22	.31
POSDEBQ20	.45	.03

Note. Exploratory Factor Analysis (EFA) extraction using Principal Axis Factoring (PAF) as the

factor extraction method and oblimin rotation. Bolded values represent double-loaded items.

Bolded items represent double loaded items. Items with an asterisk (*) are non-loaded items.

Table 5

EFA for Replication Crisis Knowledge

Items	Factor 1	Factor 2
POSRCKQ1	-.01	.84
POSRCKQ2	.62	-.02
POSRCKQ3	-.05	.40
POSRCKQ4	.44	.21
POSRCKQ5	.65	-.02
POSRCKQ6	.64	-.02
POSRCKQ7	-.56	.05

Note. Exploratory Factor Analysis (EFA) extraction using Principal Axis Factoring (PAF) as the factor extraction method and oblimin rotation.

Appendix A

Attention Check Questions

These two questions pertain to the content of the 10-minute replication crisis video. Once you've answered both questions correctly, you'll be able to proceed with the post-video questionnaire.

1. In the video, I talked about a large-scale project called the Reproducibility Project where researchers tried to excuse questionable research practices by reproducing 100 studies. The researchers running this project followed proper procedure to replicate them. How many do you think replicated?"
    a. 45%
    b. 36%
    c. 75%
    d. 26%
2. What is a reason a study might **not** replicate:
    a. *P*-Hacking
    b. Large Sample Size
    c. HARKing (Hypothesizing After the Results are Known)
    d. All are correct
    e. Only p-hacking and HARKing are correct

Appendix B

Norm and Counternorm Items (Beaudry et al., 2022)

The next set of questions relate to the following scenario:

Imagine that Deborah is a psychology researcher who has designed a study to test a specific hypothesis.

Please indicate the extent to which you agree with the following statements. Some of the questions might appear similar, but please read each statement carefully and respond to it individually.

1. Deborah used previously published research to inform her research study. Deborah should accept the findings published in journals because journals would not publish research errors.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
2. Deborah used previously published research to inform her research study. Deborah should be critical of the findings published in journals because published research can be wrong.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)

- Strongly agree (5)

3. Before collecting her data, Deborah should write down what her hypotheses are and how she plans to collect and analyse data. She should then save her plan in an online registry so others can tell what methods and analyses she will use to test her hypotheses after collecting data.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
4. Deborah should decide which data analyses are suitable to test her hypotheses only after looking at her data.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
5. Deborah should submit a plan for her study to a journal to be checked by experts (peer reviewed) *before* she collects and analyses data.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)

- Strongly agree (5)

6. Deborah should submit her study to a journal to be checked by experts (peer reviewed) only *after* she has finished collecting and analysing her data.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
7. It would be good scientific practice for Deborah to run many different analyses of her data, and report those that produce interesting findings.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
8. Deborah should report the findings of all analyses of her data that she conducts.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
9. Once Deborah has analysed her results, it would be good scientific practice for her to write her manuscript as if she predicted those results from the beginning.

- Strongly disagree (1)
- Somewhat disagree (2)
- Neither agree nor disagree (3)
- Somewhat agree (4)
- Strongly agree (5)

10. Deborah should only describe her study as a test of a hypothesis if she decided on her hypothesis and how she would test it before she started collecting data.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
11. When reporting the findings of her study, Deborah should describe how she completed the study in enough detail that another researcher could repeat her entire study without having to check any details with her – even if this means including lots of “boring” practical details in her report, or in an appendix.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
12. When reporting the findings of her study, it would be good scientific practice for Deborah to gloss over some of the practical details so she can tell a good story.

- Strongly disagree (1)
- Somewhat disagree (2)
- Neither agree nor disagree (3)
- Somewhat agree (4)
- Strongly agree (5)

13. Deborah should post a manuscript describing the findings of her study openly online as soon as it is complete, even if the manuscript has *not* yet been checked by experts (peer reviewed) and accepted for publication in a journal.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
14. Deborah should not post a manuscript describing the findings of her study online until after it has been checked by experts (peer reviewed) and accepted in a journal.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
15. Deborah should share the written materials and measures for her study openly online so that other researchers and members of the public can access and use them.
    - Strongly disagree (1)

- Somewhat disagree (2)
- Neither agree nor disagree (3)
- Somewhat agree (4)
- Strongly agree (5)

16. Deborah should keep the written materials and measures for her study protected, so that only she and her research team can access them.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
17. When Deborah published her study, she should post the anonymous responses from participants online so that anyone can access and use the responses in their own research.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
18. Deborah should keep the participants' responses from her study protected so that only she and her research team can access them.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)

- Somewhat agree (4)
- Strongly agree (5)

19. Once her study is complete, Deborah should publish her findings in a journal that is free for others to access.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)
20. Once her study is complete, Deborah should publish the findings in the most prestigious journal she can, even if that journal charged other a fee to access the report.
    - Strongly disagree (1)
    - Somewhat disagree (2)
    - Neither agree nor disagree (3)
    - Somewhat agree (4)
    - Strongly agree (5)

Appendix C

These following questions will ask which Open Science practices you engaged in during your thesis. Some questions will only be asked if you performed Open Science practices. Please answer as honestly as possible!

1. Did you complete a power analysis?
    a. Yes
    b. No
    c. Prefer not to say
2. Did you pre-register your study plans and hypotheses before conducting research?
    a. Yes
    b. No
    c. Prefer not to say
3. If you answered YES to the previous question, what platform did you use?
    a. ClinicalTrials.gov
    b. AsPredicted
    c. OSF
    d. Other. _____
    e. Prefer not to say
4. Are you sharing your data openly (i.e., making your research publicly available)?
    a. Yes
    b. No
    c. Prefer not to say
5. If you answered YES to the previous question, what platform did you use?
    a. Dryad
    b. Figshare

   c. OSF
   d. Zenodo
   e. Other. _____
   f. Prefer not to say

6. Do you have a plan for publication/dissemination regardless of results?
   a. Yes
   b. No
   c. I haven't figured that out yet
   d. Prefer not to say
7. Was practicing proper Open Science encouraged by your advisor?
   a. Yes
   b. No
   c. Prefer not to say
8. If you performed Open Science practices, did your advisor navigate you through the practices. For example, if you were unsure on how to pre-register, did your advisor help you through the process?
   a. Yes
   b. No
   c. Prefer not to say

Appendix D

Attitudes Toward Psychology (Chopik et al., 2018)

Below are some questions about your attitudes toward psychology

How much do you agree with each statement? Please answer honestly.

1. I like psychology.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree
(6) agree
(7) strongly agree

2. I think psychological research is similar to research in fields like chemistry, physics, or biology.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree
(6) agree
(7) strongly agree

3. I think psychological research is similar to research in fields like philosophy, literature, or modern languages.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree
(6) agree
(7) strongly agree

4. I think psychology is a “soft” science.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree

(6) agree
(7) strongly agree

5. I am interested in pursuing graduate school in psychology.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree
(6) agree
(7) strongly agree

6. I trust the results of studies done by psychologists.

(1) strongly disagree
(2) disagree
(3) slightly disagree
(4) neither disagree nor agree
(5) slightly agree
(6) agree
(7) strongly agree

Appendix E

Study Design Knowledge (Chopik et al., 2018)

Please rate how important you consider each aspect of study designs.

1. For a researcher, how important is choosing a sample size before running a study?

1-not important at all
2-slightly important
3-moderately important
4-very important
5-extremely important

2. It is important for a researcher to disclose all measures and experimental conditions that were included in a study.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

3. How important is it to make data publicly available so that results can be verified by other researchers?

1-not important at all
2-slightly important
3-moderately important
4-very important
5-extremely important

4. How important are decisions in data collection, analysis, and reporting in affecting how likely a researcher will find a significant effect?

1-not important at all
2-slightly important
3-moderately important
4-very important
5-extremely important

5. How important is it to report studies that “don’t work out”?

1-not important at all
2-slightly important
3-moderately important
4-very important
5-extremely important

6. How important is it that results from a psychology study are counter-intuitive (e.g., different from what you would expect)?

1-not important at all
2-slightly important
3-moderately important
4-very important
5-extremely important

Appendix F

Replication Crisis Knowledge (Chopik et al., 2018)

Please indicate your agreement with the following statements. Please answer each question honestly.

1.The field of psychology has problems replicating results.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

2. Replication of research is only a problem in the field of psychology.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

3. The incentive structure in psychological research can undermine the broader goals of science.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

4. The results from studies with low statistical power are by definition incorrect.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

5. Researchers who perform replication studies are not qualified to conduct psychological research.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

6. Studies that receive a lot of media attention are often reliable.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

7. It is important for a researcher to report all measures and experimental conditions that were included in a study.

1-disagree strongly
disagree
disagree slightly
neither disagree nor agree
agree slightly
agree
7-agree strongly

Appendix G

Demographics

Thank you for participating in this study! Before getting started, I'd like to know a bit more about you. All responses remain anonymous!

1. What is your age?

-18 to 25 years old

-26 to 35 years old

-36 to 45 years old

-46 to 55 years old

-55+ years old

2. Please indicate how you would best describe your ethnic or cultural background using the general categories presented below. Check as many as apply.

-Arab/West Asian (e.g., Armenian, Egyptian, Iranian, Lebanese, Moroccan)

-Black (e.g., African, African-American, African-Canadian, Afro-Caribbean)

-East Asian (e.g., Chinese, Japanese, Korean)

-Latinx (e.g., Brazilian, Colombian, Mexican)

-Indigenous (e.g., First Nations, Inuit, Métis)

-Pacific Islander (e.g., Fijian, Native Hawaiian, Samoan)

-South Asian (e.g., East Indian, Pakistani, Punjabi, Sri Lankan)

-South East Asian (e.g., Cambodian, Filipino, Indonesian, Laotian, Vietnamese)

-White/European (e.g., English, French, Scottish, Polish)

-If a group is not listed above that best represents your ethnic identity, please specify here: _____

3. How do you describe yourself?

-Male
-Female
-Non-binary/third gender
-Prefer to self-describe. _____
-Prefer not to say

4. What is your year in school?

-Freshman (year 1)

-Sophomore (year 2)
-Junior (year 3)
-Senior (year 4)
-Senior+ (year 5 or more)

5. Are you a psychology major? **(ONLY PSYC 1200 STUDENTS ANSWER THIS)**

-Yes

-No

6. (If they respond no) What is your university major? **(ONLY PSYC 1200 STUDENTS ANSWER THIS)**