How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users

Nicholas W. Fox[1], Nathan Honeycutt[1], & Lee Jussim[1]

[1] Rutgers University

Author Note

Department of Psychology, Rutgers University, Piscataway NJ 08854

Correspondence concerning this article should be addressed to Nicholas W. Fox, 53 Avenue E, Room 429, Piscataway NJ 08854. E-mail: nwf7@psych.rutgers.edu

Abstract

Psychology has been in crisis. Over the past 15 years many high-impact research findings have failed to replicate, calling into question their validity. Increased methodological and statistical scrutiny has led to field-wide introspection on how best to produce robust, reproducible, research. One focus has been on the use of questionable research practices. Previous estimates of the number of researchers who use questionable research practices vary widely, from over 60% to near 10%. In the current work, the authors produced three estimates of the number of American psychologists who have used questionable research practices in the last 12 months, utilizing direct, indirect, and social network measures of estimation. We estimate up to 24.40% of American psychologists have recently used at least one questionable research practice. These estimates represent the first step in generating actionable interventions to resolve the current replication crisis and increase trust in published research.

Keywords: questionable research practices, QRPs, social networks, population size estimation

Word count: X

How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users

It is the researcher's job to generate theories, test hypotheses, collect and interpret data, interpret results, and to publish findings. This is all done to learn more about the world and how it works. In the course of doing science, the researcher has many decisions to make: How many subjects will I use? How will I operationalize my variables? What is my population of interest? Should I exclude any data from the analysis?

Each decision point is a "researcher degree of freedom" (Simmons, Nelson, & Simonsohn, 2011), with the potential of introducing error and bias. Since there is a high level of ambiguity in research, these degrees of freedom can be resolved in different ways. In reviewing how researchers deal with outlying observations, Simmons et al. (2011) found different research groups made independent decisions on the best course of action. When researchers cleaned data and removed participants that responded "too fast", some defined this as 2 standard deviations below the mean response speed, some defined it as observations below 200ms, and others removed the fastest 2.5% of respondents. None of these definitions are inherently an incorrect interpretation of "too fast", which can be a problem: without clear standards in place, this type of flexible decision making can blur the lines between what decision is right, what decision produces the desired result, and what decision is most likely to help get a finding published.

There are many "degrees of freedom" that exploit the gray areas of acceptable practice that may bias research findings (John, Loewenstein, & Prelec, 2012; Wicherts et al., 2016). Some examples include trying different ways to score the chosen primary dependent variable, and deciding how to deal with outliers in an ad hoc manner. Ten of these behaviors have been collectively called "questionable research practices" (QRPs) and are typically defined as behaviors during data collection, analysis, and reporting that have the potential to increase

false-positive findings in the literature. While there are many examples of other behaviors that could be considered questionable, these ten stand out as being familiar to most researchers and having been investigated previously (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Fiedler & Schwarz, 2016; John et al., 2012). For this study, nine of the ten QRPs were considered (Table~**??**). We did not include "fabricated data" (QRP-10) as a questionable research practice as the authors consider this a fraudulent, not questionable, behavior. Not only can QRP use increase the number of false-positive findings (e.g., taking a "non-significant" result and pushing it over a threshold into being "significant"), but using multiple QRPs can also influence the reported effect size of a given finding due to sampling bias and low power (Button et al., 2013). Thus, QRP use could lead to field-wide interpretations that are not warranted by the data.

Prevalence of questionable research practices

Consider one of the most basic questions to ask about the current replication crisis: How many people are contributing to it? John et al. (2012) found 63% of psychologists admitted to publishing work without all the dependent measure included (at some point in their academic career). As articulated by Simmons et al. (2011), this is highly problematic because increasing the number of dependent variables is correlated with an increase in the probability of finding a significant result. Without reporting all dependent measures, readers are left with a false impression of the rarity or truthfulness of the reported findings. But, this estimate from John et al. (2012) was contested by Fiedler and Schwarz (2016). In their conceptual replication that used differently worded questions, used a different conceptualization of "prevalence", and tested a German (as opposed to an American) cohort of psychologists, Fiedler and Schwarz (2016) found less than 10% prevalence of the same questionable practice (omitting dependent variables).Furthermore, Agnoli et al. (2017) recently replicated the original John et al. (2012) study in an Italian cohort of psychologists,

and found similarly high levels of QRP use (47.9% of respondents had omitted dependent variables). Concequently, there is no current consensus on the prevalence of QRP use in psychology, nor any indication of how these may be related to the current replication crisis in the field.

Given the inconsistencies in assessing the prevalence of QRP use, the present work sought to expand on this existing literature in several ways. First, we investigated current QRP users, operationalized as a person who has used at least one of nine QRPs "in the past 12 months". This puts QRP use into the timeframe of the current replication crisis. Second, it addressed the larger issue of "prevalence", by defining behaviors performed within a specified time period. Previous work estimating QRP prevalence has done so over career-long timespans or via estimating frequency of QRP use, both providing limited insight on the current issues in the field.

A third unique contribution of the present research is that it assessed prevalence of QRP use with several starkly different methodologies. One was a direct estimate tightly based on prior research (Agnoli et al., 2017; Fiedler & Schwarz, 2016; John et al., 2012): We simply asked researchers to report their own QRP use.

However, two other methods were increasingly different from this straightforward assessment. One used the unmatched count technique, an indirect estimate aimed at reducing social desirability response bias (CITATION; see Method for details). A third generated an indirect estimate of QRP use by using social network information from the general population of psychologists (Jing, Qu, Yu, Wang, & Cui, 2014; M. J. Salganik et al., 2011; Zhang et al., 2010; Zheng, Salganik, & Gelman, 2006). Neither the unmatched count technique nor these social network methods required participants to identify as belonging to a potentially stigmatized group (QRP users), thereby reducing the risk of socially desirable response bias compared to more traditional direct estimates. While network methods were expected to provide insights into QRP use prevalence, they have yet to be used in

psychology. Thus, this work produced three estimates of QRP use prevalence.

METHOD

The work detailed in this manuscript was preregistered on May 15th, 2017. The preregistration can be found at www.osf.io/xu25n and is detailed in the supplemental materials.

Sample

The frame population was tenured or tenure-track faculty associated with a psychology department at a PhD-granting institution in the United States. The population contained 7,101 individuals as of June 2017. All 7,101 members of this population were contacted via email and asked to participate. Of the 7,101 email invitations sent, 214 emails bounced (3.01%). We collected 613 full responses (8.63% full response rate), and 296 partial responses. Only full responses are used in the following estimations. Additionally, 26 participant responses were removed for either being marked complete erroneously or due to breaking estimate-specific criteria. There was no compensation offered for participation. 299 (48.78%) participants identified as female, 279 (45.51%) identified as male, and 19 (3.10%) chose not to identify their gender. 131 (21.37%) participants identified as an Assistant Professor, 141 (23.00%) as Associate Professor, and 208 (33.93%) as Full Professor. 113 participants identified as tenure or tenure-track, but did not disclose their tenure level.

Data Sources

Data was collected using three surveys (as opposed to the two proposed in the preregistration, see supplemental materials), designed and distributed using Qualtrics survey software (CITATION). Each survey asked questions designed to estimate the total social

network size of the participant, as well as demographic questions. Surveys 1 and 2 each contained questions appropriate for the unmatched count technique (UCT). Survey 3 contained, instead of the UCT, our direct estimate measure and questions used to determine transmission of QRP-identity information within an individual's social network.

Participants were contacted by email and provided a link to one of the Qualtrics surveys. Upon agreeing to the informed consent, participants completed the aforementioned measures. To ensure the highest number of participants in our game of contacts procedure (see methods), half of the total population were asked to participate in Survey 3, which contained our direct estimate question. Thus, 3551 were solicited for Survey 3. The remaining 3550 psychologists contacted were asked to participate in our unmatched count estimate (Survey 1 or Survey 2), with 1775 randomized into the innocuous list group, and 1775 randomized into the concealable list group. Participants had seven days to complete the survey upon starting. Since data was collected between September 2017 and December 2017, questions framed "in the past 12 months" bound actual QRP use between September 2016 and December 2017, a time frame of 15 months. Therefore, estimates of "current QRP use" are based on the number of psychologists who have used at least one QRP in this time frame.

All surveys included the definition of "Questionable Research Practices (QRPs)". This definition included the list of behaviors previously defined in the literature as QRPs (see Table), but omitting "fabricating data" for reasons addressed earlier. The definition of QRP was made available on each relevant question with a mouse rollover that was first demonstrated with the initial definition. For the full text of our definition used, see supplemental materials.

Additionally, QRP use (as defined in the present work) is constrained to behaviors performed "in the past 12 months". Although some have found instances of underreporting when using a 12 month recall (Connelly & Brown, 1995; Landen & Hendricks, 1995), this time frame is used frequently to measure current behavior in major national data collection

surveys such as the National Health Interview Survey (NHIS) (United States Census Bureau, 2018) and the National Survey on Drug Use and Health (Ahrnsbrak, Bose, Hedden, Lipari, & Park-Lee, 2017).

Measures

Direct Estimate. The direct estimate involved asking members of the target population whether they have used at least one QRP in the past 12 months, and is calculated as follows:

$$\rho = \frac{c}{n} \tag{1}$$

where ρ is the proportion estimate of people who have used at least one QRP in the past 12 months, c is the number of participants indicating they have used a QRP in the past 12 months, and n is the total number of participant responses.

Unmatched Count Technique. The unmatched count technique (UCT) is an indirect way of measuring the base rates of concealable and potentially stigmatized identities (Gervais & Najle, 2017; Wolter & Laier, 2014). In this estimate, two groups of participants are given a list of innocuous items that could apply to them (e.g., I own a dishwasher; I exercise regularly). The list of items for both groups is the same except for one additional item that one group receives and the other does not. This extra item asks about the concealable identity (e.g., I own a dishwasher; I exercise regularly; I smoke crack cocaine [examples from (Gervais & Najle, 2017)]). See Table~**??** for the full list of items used. Participants are asked to count and report the number of items in the list that apply to them. At no point does a participant identify themselves with any particular list item, only the total number of applicable items. The proportion of participants that identify with the stigmatized identity is calculated as:

$$\rho = \frac{\sum x_y^s}{n^s} - \frac{\sum x_y^i}{n^i} \tag{2}$$

where ρ is the proportion estimate of people who have used at least one QRP in the past 12 months, x_y^s is the number of reported items for participant y in the stigma list group s, n^s is the total number of participant responses in group s, x_y^i is the number of reported items for participant y in the innocuous list group i, and n^i is the total number of participants in group i.

Network Methods. Network methods estimate population sizes using information about the personal networks (referred to as "ego networks" in this literature, i.e., (McCormick, Salganik, & Zheng, 2010)) of respondents, based on the assumption that personal networks are, on average, representative of the population (M. J. Salganik et al., 2011). Each participant's social network provides a sample of the general population, and by collecting network data on many participants, those accumulated social networks provide access to the larger population.

Network Scale-Up Method (NSUM).

Participants were asked about how many people they "know" in the frame population. In this study, "know" was defined as: they know you by face or by name, you know them by face or by name, you could contact the person if you wanted to, and you've been in contact in the past two years (H. R. Bernard et al., 2010). Participants were then asked a series of questions to estimate the total size of their social network, and the number of people they know who have used at least one QRP in the past 12 months. Together, the network scale-up can be used to estimate the proportion of QRP users, and was calculated as follows:

$$\rho = \frac{\sum y_i}{\sum d_i} \tag{3}$$

where ρ is the proportion estimate of people who have used at least one QRP in the past 12 months, y_i is the number of people known in the target group y by participant i, and d_i is the estimated total number of people known d by participant i within the frame population (see (Killworth, McCarty, Bernard, Shelley, & Johnsen, 1998) for more on estimating d). Equation 3 makes two assumptions: that members of the general frame population know all identity information about all members of their ego networks, and that QRP users have the same size social networks as the general frame population.

Generalized Network Scale-Up Method (GNSUM).

Since QRP use is concealable and potentially stigmatizing, the assumptions made for Equation 3 not hold. For that reason, data was collected from self-identifying QRP users to estimate how QRP-use identity information transmits through ego networks. This estimate is called the transmission rate, or tau (τ). This data was collected using the game of contacts method (Salganik et al., 2012), described below.

To estimate the QRP use identity transmission rate, τ, we performed the game of contacts with participants who self-identified as using at least one QRP in the past 12 months. Briefly, this method has participants (called egos in network terminology) answer a set of questions about what they know about the QRP use of several others (called alters) in their social network, and what those alters know about the participant's QRP use. The questions are semi-graphical and responses are recorded on a digital 2x2 grid, representing the four possible ways information can flow through a given ego-alter relationship. The transmission rate is then calculated as:

$$\tau = \frac{\sum w_i}{\sum x_i} \tag{4}$$

where w_i is the number of alters that know the ego is a member of the target population, and x_i is the total number of alters generated by the ego. This produced a value

between 0 and 1. For a full description of the game of contacts, see Salganik et al. (2012).

This study utilized a digital distribution of the game of contacts. This method is typically performed in a face-to-face interview setting with the participant (Salganik2012b). Due to the distributed nature of our frame population, this was not feasible. Instead, participants were presented with the game of contacts via Qualtrics. These questions were pretested with several academics not within the frame population. A comparison between an in-person and digital game of contacts has been pre-registered by the authors (https://osf.io/yf4xc/) for future study.

Additionally, to relax the assumption of equal social network sizes between the general frame population and QRP users, a popularity ratio (delta, δ) was calculated by dividing the average network size of QRP users by the average network size of the general frame population.

Together, τ and δ adjust the network scale-up estimate in Equation 3 into the generalized network scale-up as follows:

$$\rho = \frac{\sum y_i}{\sum d_i} * \frac{1}{\tau} * \frac{1}{\delta} \tag{5}$$

where ρ is the proportion estimate of people who have used at least one QRP in the past 12 months, $\frac{\sum y_i}{\sum d_i}$ is the network scale-up estimate (equation 3), τ is the transmission rate, and δ is the popularity ratio. All network scale-up results are calculated using Equation 5, incorporating τ and δ.

Results

The three estimates of recent QRP use in the frame population of American tenured or tenure-track faculty are summarized in Figure 1, and described in detail below.

Direct Estimate

To ensure the highest number of participants in our game of contacts, half of the total population were asked to participate in Survey 3, which contained our direct estimate question. Thus, 3,551 psychologists were solicited, and we received 308 responses to Survey 3 able to be analyzed. Of the 308 participants, 56 indicated they had used at least one QRP in the past 12 months. Using Equation 1, we calculated QRP prevalence to be 18.18% (bootstrapped 95% confidence interval [13.96%, 22.40%]). This corresponds to an estimated 1,291 American psychologists currently using QRPs.

It is possible this estimate underestimates the true number of psychologists using QRPs. For one, social desirability may lead some scientists who have used QRPs to be unwilling to admit it. This estimate is only generated by those participants willing to reveal their identity as a QRP user. Given the somewhat critical social environment for QRP users (Fiske, 2016; Teixeira da Silva, 2018), it is reasonable to believe some participants withheld their identity when we asked directly. The following indirect estimation methods sought to mitigate this social desirability bias.

Unmatched Count Technique

The remaining 3,550 psychologists contacted were asked to participate in our unmatched count estimate with 1,775 randomized into the innocuous list condition, and 1,775 randomized into the sensitive list condition.

The average number of list items corresponding to participants in the innocuous list condition was 4.28. The average number of list items corresponding to participants in the sensitive list condition was 4.39. Using Equation 2, we calculated QRP user prevalence to be 10.46% [-20.19%, 22.40%]. This corresponds to an estimated 743 American psychologists currently using QRPs.

It was unexpected that the calculated UCT estimate would be lower than our direct estimate. Typically, due to reducing response bias, UCT estimates are larger than direct estimates when the behavior or identity in question is concealable and potentially stigmatized (Gervais & Najle, 2017; Starosta & Earleywine, 2014; Wolter & Laier, 2014). Given the bootstrapped 95% confidence interval crosses zero, it is likely the relatively low number of participants in our UCT (n = 279) led this calculation to be overly sensitive to individual responses, and as such, we do not consider this estimate to be valid or accurate.

Generalized Network Scale-Up Estimate

All participants who were randomized into the UCT estimate were also asked to answer questions about their social networks, and to estimate how many researchers they know who have used at least one QRP in the past 12 months. Participants who were randomized into the direct estimate and who self-identified as a QRP user in that estimate were also asked to answer questions about their social network and to participate in the game of contacts method. Participants in the direct estimate who did not self-identify as a QRP user were asked questions about their social network as well, but were not asked how many researchers they know who have used at least one QRP in the past 12 months. Therefore, we collected social network responses from 531 participants from the general frame population (to be used in estimating δ), 56 responses from participants who self-identified as QRP users who also completed the game of contacts (to be used in estimating τ), and 279 responses from participants who estimated the number of researchers they know who have used at least one QRP in the past 12 months.

These 279 identified a sum total of 664 QRP users, and know a sum total of 46,828 researchers. Given the total frame population is 7,101, we are fairly confident all or nearly all members were identified at least once by our participants. Using the network scale-up in Equation 3, this generates an estimate of 1.42% [0.85%, 2.14%]. This estimate serves as the

base starting point our key network estimate, the Generalized Network Scale-Up Estimator, detailed below.

Equation 5 relaxes the assumptions of equal network size and total information transmission by incorporating τ and δ. Using the 531 responses from the general population and the 56 responses from the participants who indicated using a QRP in the past 12 months, we estimate δ as 0.97. Using the game of contacts, we estimate τ as 0.06. Using Equation 5, we estimate QRP user prevalence to be 24.40% [10.93%, 58.74%]. This corresponds to an estimated 1,733 American psychologists currently using QRPs.

To assess the accuracy of participants in estimating the size of this unknown group (QRP users), we generated additional estimates of 24 populations of known size; the number of psychologists with particular first names (the number of psychologists named David, named Janet, etc). The 24 names were gender balanced and represented common, uncommon, and rare names that exist within the census of the frame population. The size estimates of these populations of known size can be seen in Figure 2. The estimates made by our participants of the size of these 24 populations seem reasonable and closely mirror the actual prevalence of these groups. The correlation between our participant's estimate of those group sizes and the actual group sizes is r = 0.91. The fact that the same estimator in the same group of participants can generate reasonable estimates for populations of known size is encouraging evidence of the accuracy of our estimate of the number of recent QRP users utilizing the generalized network scale-up estimate.

DISCUSSION

Because of inconsistencies in previous research, this study generated three estimates of current QRP use, using three independent estimating procedures. Depending on the estimator used, we estimate 18.18% to 24.40% of American psychologists currently use

questionable research practices. Our unmatched count estimate produced an estimate of 10.46%, though this may not be a valid or accurate estimate.

To the best of our knowledge, this is the first report of the prevalence of QRP users in a proximal timespan. As such, it is difficult to draw conclusions about the magnitude of our estimates when compared to previous estimates.

Compared to John et al. (2012) and Agnoli et al. (2017), we estimate lower rates of questionable research practices. Compared to Fiedler and Schwarz (2016), however, we estimate higher rates of these practices. Our definition of "questionable research practices" were the same ones used in John et al. (2012) and Agnoli et al. (2017), but was restricted to a timespan of only 15 months, so it is reasonable that our estimates would be lower than those with an unrestricted timeframe of QRP use. Since we used those same QRP definitions, is also reasonable that our estimates would be higher than those described by Fiedler and Schwarz (2016), who changed the definitions of each QRP.

This is also the first report to use the generalized network-scale up estimators to investigate the prevalence of QRP users in psychology. Direct estimates rely on an individual's willingness to participate and their willingness to honestly share their identity as a QRP user. Bias in either of these dimensions can distort a direct estimate.

Social network methods, on the other hand, enable researchers to better understand the social processes at work that produce an environment where members vary in their identity and the information they share with others (Zheng et al., 2006). In the process of producing a population size estimate for current QRP users, we also reported the first estimate of the social transmission of this QRP identity - tau (τ).

Our reported estimate ofτ (0.06 or 6.02%) means very few individuals who use QRPs share this identity with others, choosing instead to conceal this identity from their peers. Because of this, it is important to shift the current conversation on QRPs away from isolated

behaviors (i.e., prevalence of rounding down p values) and towards the usuers of these behaviors (i.e., prevalence of people who round down p values). Re-framing QRP prevalence away from the behavior and towards the users brings our field-wide problems more into scope with existing literature on concealable identities and stigma. For example, much work has been done focusing on how increasing stigma inadvertently locks individuals into detrimental behaviors (Bayer & Colgrove, 2002; Stuber, Galea, & Link, 2009), and how revealing a concealed identity can increase well-being by reducing the stress of being exposed (Chaudoir & Quinn, 2010). Framing QRP use in terms of the individual may help the field reduce QRP use by increasing awareness of the effects of stigma and support.

Implications

These estimates serve as a baseline to measure the effectiveness of current initiatives, as well as a foundation for new ones. While much work is being done to grow support for interventions such as pre-registration (E.-J. Wagenmakers & Dutilh, 2016) and Registered Reports (D. Chambers, Feredoes, D. Muthukumaraswamy, & J. Etchells, 2014), it is currently unknown what quantitative effect these are having on curbing behaviors associated with inflated Type I error such as QRPs. By performing follow-up estimates at future time points, the field can use the baseline estimates presented here to measure the effectiveness of these programs at reducing QRP use.

Limitations & Future Directions

Our unmatched count estimate was lower than our direct estimate and had a confidence interval that included zero, which we did not expect. This estimate is computed as a difference between averages, which means each is sensitive to the number of participants. Our innocuous item group had 130, and our concealable item group had 149. As

denominators in computing the means needed for Equation 2, these are low. For example, if one additional participant in the innocuous list condition responded by itentifying with 6 of 9 items (1 standard deviation above the mean), our UCT estimate would change from 10.46% to 9.15%. This is reflected in the bootstrapped confidence interval crossing zero, indicative of an unstable difference between the innocuous and concealable list groups. Future work using the UCT would benefit from larger sample sizes, as demonstrated in Gervais and Najle (2017).

As noted in the introduction, QRPs exist in a grey area of accepted scientific practice. Therefore, it is difficult to interpret the severity of QRP use. This difficulty, along with the high variability among previous estimates of QRP prevalence, has led to a number of different conclusions. Some have concluded that the problems are overstated (Fanelli, 2018), while others argue QRP use presents a real threat to the viability of several scientific fields, such as education and political science (Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Although our work moves the field forward in understanding the prevalence of those that use these behaviors, it provides less guidance on the severity of the consequences of QRP use on the whole.

Science is a globally distributed network, and as such, can be difficult to study. Our reported estimates were limited to American psychologists, though we know that these issues are not restricted solely to the United States (Agnoli et al., 2017; Fiedler & Schwarz, 2016; Forsberg et al., 2018). Future studies estimating the prevalence of QRP use in other countries will be an important next step, as will investigating the use of QRPs in other scientific fields. Some of this work has already started through the Horizon 2020 framework in the European Union (Forsberg et al., 2018), though more innovative work will be required to better understand the scope of the problems faced.

Conclusion

By directly asking participants about their use of QRPs, we estimate 18.18% have used at least one QRP in the past 12 months. The generalized network scale up estimate is 24.40%, which corresponds to between 1,291 and 1,733 American psychologists. Although some have argued the narrative of the "replication crisis" is overblown (Fanelli, 2018), the current work illustrates how common QRP use is. Although many have called for changes in statistical inference practices to mitigate false-positive findings (Benjamin et al., 2017; Lakens et al., 2018), it is important that we as a field also focus on disincentivizing the use of questionable research practices (and other behavioral degrees of freedom) among our peers and coworkers for the betterment of our science.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE*, *12*(3), 1–17. doi:10.1371/journal.pone.0172792

Ahrnsbrak, R., Bose, J., Hedden, S., Lipari, R., & Park-Lee, E. (2017). Key Substance Use and Mental Health Indicators in the United States: Results from the 2016 National Survey on Drug Use and Health. *Substance Abuse and Mental Health Services Administration*, *7*(1), 877–726. doi:10.1016/j.drugalcdep.2016.10.042

Bayer, R., & Colgrove, J. (2002). Science, politics, and ideology in the campaign against environmental tobacco smoke. *American Journal of Public Health*, *92*(6), 949–954. doi:10.2105/AJPH.92.6.949

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., & Johnson, V. E. (2017). Redefine Statistical Significance. *PsyArxiv*, (July 22), 1–18. doi:10.17605/OSF.IO/MKY9J

Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., ... Stroup, D. F. (2010). Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, *86 Suppl 2*, ii11–5. doi:10.1136/sti.2010.044446

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's Threat to Organizational Research: Evidence From Primary and Meta-Analytic Sources. *Personnel Psychology*, *69*(3), 709–750. doi:10.1111/peps.12111

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*(5), 365–376. doi:10.1038/nrn3475

Chaudoir, S. R., & Quinn, D. M. (2010). Revealing Concealable Stigmatized Identities: The Impact of Disclosure Motivations and Positive First-Disclosure Experiences on Fear of Disclosure and Well-Being. *Journal of Social Issues*, *66*(3), 570–584. doi:10.1111/j.1540-4560.2010.01663.x

Connelly, N. A., & Brown, T. L. (1995). Use of Angler Diaries to Examine Biases Associated with 12-Month Recall on Mail Questionnaires. *Transactions of the American Fisheries Society*, *124*(3), 413–422. doi:10.1577/1548-8659(1995)124<0413:uoadte>2.3.co;2

D. Chambers, C., Feredoes, E., D. Muthukumaraswamy, S., & J. Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17. doi:10.3934/Neuroscience.2014.1.4

Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America2*, *in press*, 1–4. doi:10.1073/pnas.1708272114

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. doi:10.1177/1948550615612150

Fiske, S. T. (2016). Mob Rule or Wisdom of Crowds. *APS Observer*.

Forsberg, E. M., Anthun, F. O., Bailey, S., Birchley, G., Bout, H., Casonato, C., … Zöller, M. (2018). Working with Research Integrity—Guidance for Research Performing Organisations: The Bonn PRINTEGER Statement. *Science and Engineering Ethics*, 1–12. doi:10.1007/s11948-018-0034-4

Gervais, W. M., & Najle, M. B. (2017). How many atheists are there? *Social Psychological*

and Personality Science, 1948550617707015.

Jing, L., Qu, C., Yu, H., Wang, T., & Cui, Y. (2014). Estimating the sizes of populations at high risk for HIV: A comparison study. *PLoS ONE*, *9*(4), 1–6. doi:10.1371/journal.pone.0095601

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

Killworth, P., McCarty, C., Bernard, H. R., Shelley, G. A., & Johnsen, E. C. (1998). Estimation of Seroprevalence, Rape, and Homelessness in the U.S. Using a Social Network Approach. *Evaluation Review*, *22*, 289–308.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., J Apps, M. A., Argamon, S. E., ... Lino de Oliveira, C. (2018). Justify Your Alpha: A Response to " Redefine Statistical Significance". *Nature Human Behavior*, *2*, 168–171.

Landen, D. D., & Hendricks, S. (1995). Effect of recall on reporting of at-work injuries. *Public Health Reports (Washington, D.C. : 1974)*, *110*(3), 350–4. doi:10.1016/s0022-4375(97)90342-x

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). How many people do you know? Efficiently estimating personal network size. *J Am Stat Assoc.*, *105*(489), 59–70. doi:10.1016/j.immuni.2010.12.017.Two-stage

Salganik, M. J., Fazito, D., Bertoni, N., Abdo, A. H., Mello, M. B., & Bastos, F. I. (2011). Assessing network scale-up estimates for groups most at risk of HIV/AIDS: Evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. *American*

Journal of Epidemiology, *174*(10), 1190–1196. doi:10.1093/aje/kwr246

Salganik, M. J., Mello, M., Abdo, A., Bertoni, N., Fazito, D., & Bastos, F. (2012). The Game of Contacts: Estimating the Social Visibility of Groups, *100*(2), 130–134. doi:10.1016/j.pestbp.2011.02.012.Investigations

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Starosta, A. J., & Earleywine, M. (2014). Assessing base rates of sexual behavior using the unmatched count technique. *Health Psychology and Behavioral Medicine*, *2*(1), 198–210. doi:10.1080/21642850.2014.886957

Stuber, J., Galea, S., & Link, B. G. (2009). Stigma and Smoking: The Consequences of Our Good Intentions. *Social Service Review*, *83*(4), 585–609. doi:10.1086/650349

Teixeira da Silva, J. A. (2018). Freedom of Speech and Public Shaming by the Science Watchdogs. *Journal of Advocacy, Research, and Education*, *5*(1).

United States Census Bureau. (2018). *National health Interview Survey: CAPI Manual for NHIS Field Representative* (No. January). Retrieved from ftp://ftp.cdc.gov/pub/health{_}statistics/nchs/Survey{_}Questionnaires/NHIS/2006/frmanual.p

Wagenmakers, E.-J., & Dutilh, G. (2016). Seven Selfish Reasons for Preregistration. Retrieved from https://www.psychologicalscience.org/observer/seven-selfish-reasons-for-preregistration/comment-page-1

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Aert, R. C. van, & Assen, M. A. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology*,

7(NOV), 1–12. doi:10.3389/fpsyg.2016.01832

Wolter, F., & Laier, B. (2014). The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency. *Survey Research Methods*, *8*(3), 153–168.

Zhang, T. Y., Hellstrom, I. C., Bagot, R. C., Wen, X., Diorio, J., & Meaney, M. J. (2010). Maternal care and DNA methylation of a glutamic acid decarboxylase 1 promoter in rat hippocampus. *J Neurosci*, *30*(39), 13130–13137. doi:10.1523/JNEUROSCI.1039-10.2010

Zheng, T., Salganik, M. J., & Gelman, A. (2006). How Many People Do You Know in Prison? *Journal of the American Statistical Association*, *101*(474), 409–423. doi:10.1198/016214505000001168

Table 1

Questionable Research Practices of interest with examples.

Questionable Research Practice
Failing to report all of a study's dependent measures
Collecting more data after looking to see if the results were significant
Failing to report all of a study's conditions
Stopping data collection earlier than planned because one found the result one was looking for
Rounding off p values to achieve significance
Selectively reporting studies that 'worked'
Deciding whether to exclude observations after seeing the effect of doing so on the results
Reporting unexpected findings as being predicted from the start (a.k.a Hypothesizing After the Results
Reporting results are unaffected by demographics when actually unsure or not tested

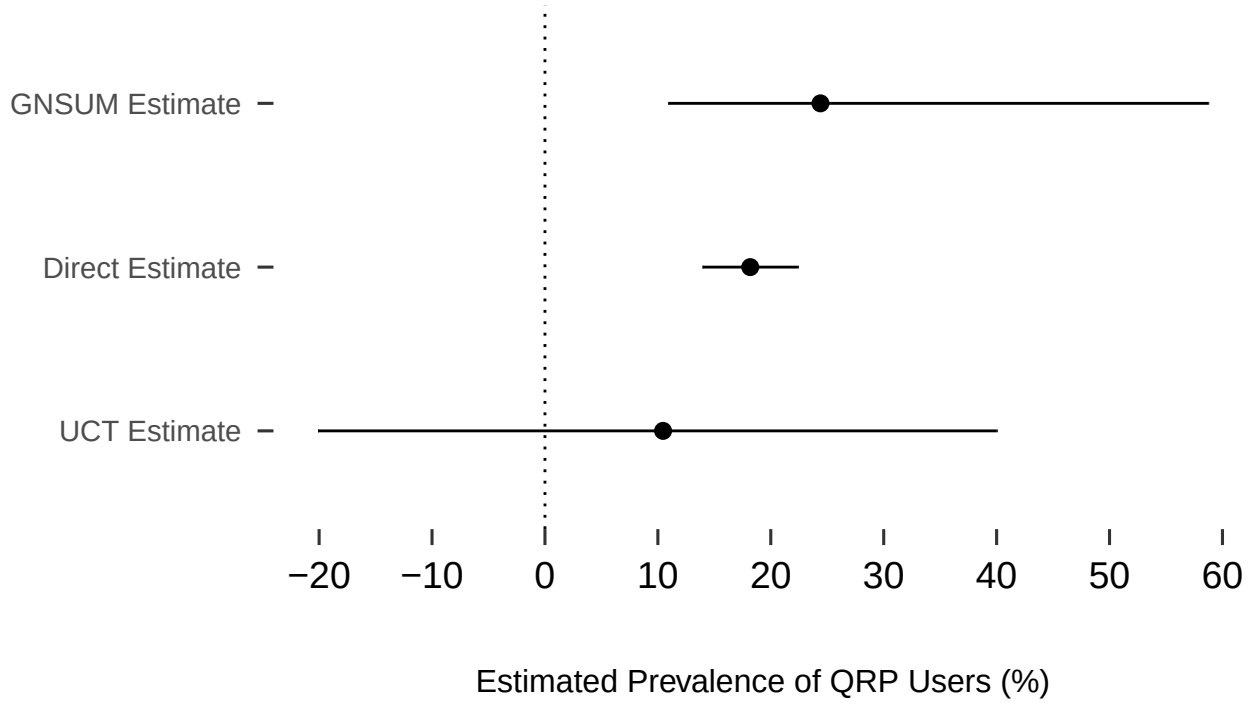

Figure 1. Estimates of the current prevalence of users of questionable research practices using three different estimators; the Generalized Network Scale-Up Method (GNSUM), the Direct Estimate, and the Unmatched Count Technique (UCT). Point estimates with 95% bootstrapped confidence intervals.

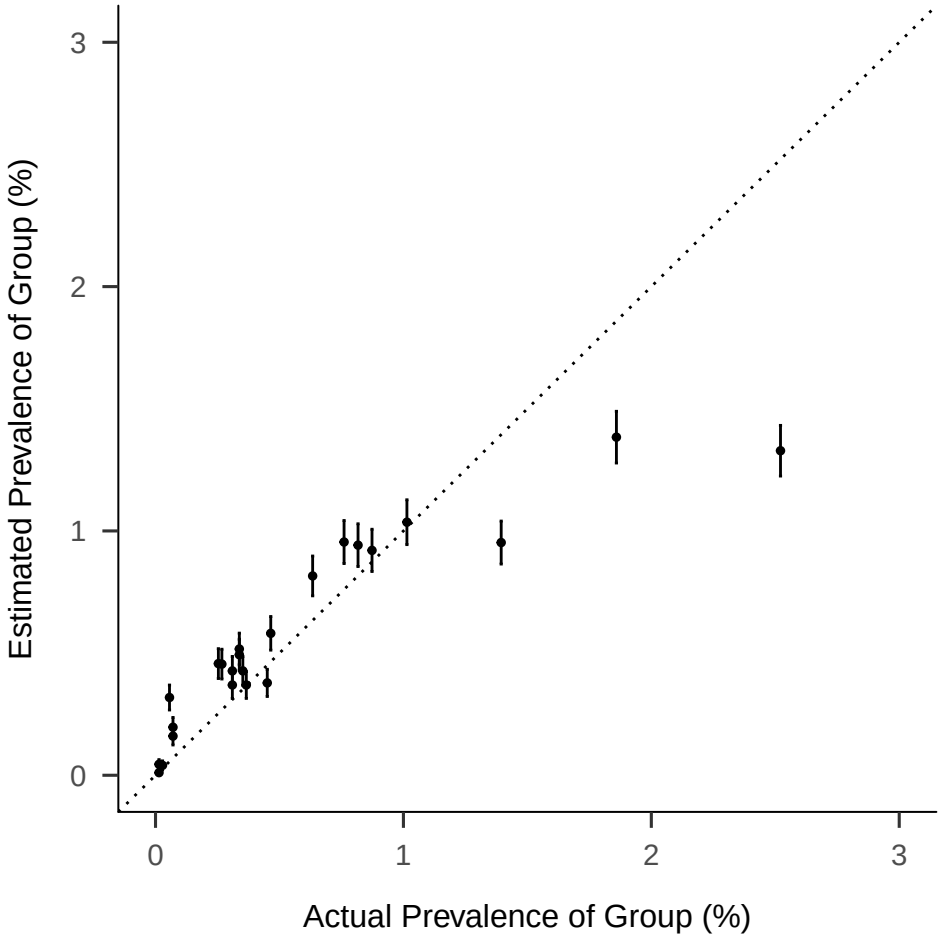

Figure 2. Validation of Network Scale-Up Estimates using 24 groups of known size. Each point represents one group, with 95% confidence intervals. Dotted line represents when estimated group prevalence equals actual group prevalence. Correlation between estimated and actual group prevalence, r = 0.91.