General Article

Advances in Methods and
Practices in Psychological Science
2019, Vol. 2(3) 214–227

DOI: 10.1177/2515245919848684
www.psychologicalscience.org/AMPPS

Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets

Sara J. Weston[1], Stuart J. Ritchie[2], Julia M. Rohrer[3,4,5], and Andrew K. Przybylski[6,7]

[1]Department of Psychology, University of Oregon; [2]Social, Genetic and Developmental Psychiatry Centre, King's College London; [3]Department of Psychology, University of Leipzig; [4]International Max Planck Research School on the Life Course, Max Planck Institute for Human Development; [5]German Institute for Economic Research (DIW Berlin), Berlin, Germany; [6]Oxford Internet Institute, University of Oxford; and [7]Department of Experimental Psychology, University of Oxford

Abstract

Secondary data analysis, or the analysis of preexisting data, provides a powerful tool for the resourceful psychological scientist. Never has this been more true than now, when technological advances enable both sharing data across labs and continents and mining large sources of preexisting data. However, secondary data analysis is easily overlooked as a key domain for developing new open-science practices or improving analytic methods for robust data analysis. In this article, we provide researchers with the knowledge necessary to incorporate secondary data analysis into their methodological toolbox. We explain that secondary data analysis can be used for either exploratory or confirmatory work, and can be either correlational or experimental, and we highlight the advantages and disadvantages of this type of research. We describe how transparency-enhancing practices can improve and alter interpretations of results from secondary data analysis and discuss approaches that can be used to improve the robustness of reported results. We close by suggesting ways in which scientific subfields and institutions could address and improve the use of secondary data analysis.

Keywords

p-hacking, bias, panel design, file drawer, reproducibility, preregistration, secondary analysis, preexisting data, transparency

Received 7/6/18; Revision accepted 4/16/19

Never before has so much human data been so widely available to researchers. Online storage platforms for academic scientists, such as Harvard Dataverse (https://dataverse.harvard.edu/) and the Open Science Framework (https://osf.io), make sharing data across labs, countries, and continents instantaneous at no cost. Government-funded data-collection initiatives organize and track individuals at an enormous scale. With the rise of social media and smartphone technology, behavioral scientists have a wide range of trace data available to analyze and combine with a rich array of data sets. However, despite this wealth of data, conversations regarding data analysis and modeling in psychology often start with the assumption that researchers collect new data for each research question they ask.

Certainly, a great many principles of *primary* data analysis (i.e., analysis of newly collected data) are still relevant, applicable, and important when preexisting data are analyzed. Nevertheless, use of preexisting data brings with it new concerns—for example, various biases and a lack of experimental control—that warrant careful consideration. On the other hand, the benefits of using preexisting data are often overlooked. In this article, we describe the analysis of preexisting data, often called *secondary* data analysis, and outline its

Corresponding Author:
Sara J. Weston, Department of Psychology, University of Oregon, 1451 Onyx St., Eugene, OR 97403
E-mail: weston.sara@gmail.com

value to psychological researchers. We also discuss the potential pitfalls of secondary data analysis, especially in view of recent advances in open science and transparency. We end with recommendations for increasing the transparency of secondary data analysis and improving the robustness of the reported results obtained, including some ideas regarding preregistration. We have written this article for scientists who are interested in adding secondary data analysis to their methodological toolbox, and for anyone who wishes to use preexisting data fruitfully and responsibly.

What Is Secondary Data Analysis?

We consider *preexisting data* to be any data that exist before researchers formulate their research hypothesis. Preexisting data can take many forms. Here, we focus on two: large-scale survey data and single-lab data.

Large-scale survey studies routinely assess a broad array of variables, often from national, representative samples and using multiple waves of assessment. Such large-scale survey studies are often formed to track changes in the attitudes, health, or economics of a population over time; consequently, they tend to be larger than single-lab studies, in terms of the number of participants sampled, the number and scope of variables assessed, and the number of members on the research team. Many panel studies—such as the German Socioeconomic Panel Study (Wagner, Frick, & Schupp, 2007), the British Household Panel Study (University of Essex, Institute for Social and Economic Research, 2018, and the National Longitudinal Study of Youth 1979 (U.S. Bureau of Labor Statistics, n.d.)[1]—are funded by governments or other large organizations and have their data made publicly available, or available upon registration.

Preexisting data do not have to be collected on a large scale. When running studies, research labs often choose to collect data that are not directly relevant to the primary research question. Alternatively, after analysis or publication of a study, researchers may think of a different question that the previously collected data may be able to answer. In both of these cases, we consider these data collected in smaller-scale lab studies to be preexisting data. Thus, the process of generating and sharing data for use by other researchers need not be left to research councils and national governments. Given the potentially limited sample size of these smaller-scale investigations, considerations of statistical power cannot be ignored when analyzing their data. Single-lab studies may resemble panel studies in that participants may be tracked over time and a variety of constructs may be measured repeatedly.

Preexisting data can take other forms as well. One of the fastest growing areas of research is focused on "big data," or data collected through the use of modern technologies, including the Internet and smartphones (Hashem et al., 2015; Kosinski, Stillwell, & Graepel, 2013). Often these kinds of social or medical data are collected without a primary research question in mind and may later be mined by researchers. We believe the claims regarding, and recommendations for, the use of preexisting data extend to analyses using big data. We consider *secondary data analysis* to be the analysis of any preexisting data.[2]

Psychologists often think about research in terms of two modes: exploratory (i.e., theory building) and confirmatory (i.e., theory testing; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Exploratory research is a common focus for secondary data analysis and is one of its great strengths. Because preexisting data sets often contain many—even many thousands—of variables, researchers have the flexibility to explore many relationships between constructs. Researchers may run exploratory analyses of preexisting data sets without wasting valuable time or financial resources. If they find evidence of a relation, they can choose to invest in another study to confirm it; if they find little evidence, they may decide it would be a waste of resources to collect new data.

On the other hand, it is also possible to use preexisting data to test theories in a confirmatory fashion. However, this endeavor comes with an important caveat: Many commonly applied statistical tests were developed under specific assumptions. For example, null-hypothesis significance testing assumes that the statistical test is chosen prior to data collection; this is part of what makes data peeking so problematic in research (Armitage, McPherson, & Rowe, 1969; Munafò et al., 2017). Consequently, researchers conducting secondary data analyses that might help confirm a theory must take extra steps to ensure the robustness of their results. We describe some of these possible steps—which, it should be noted, are not mutually exclusive—in Table 1.

Psychological research can also be categorized as correlational or experimental, and secondary data analysis can be either. It is true that correlational work makes up the bulk of secondary data analysis, given that much of such analysis uses data from panel studies and other surveys (see Rohrer, 2018, for a discussion of causality in psychological research). However, in the case of single-lab studies, experimental work might also fall under the umbrella of secondary data analysis. For example, data from a study designed to assess the effectiveness of an intervention on academic performance might be reanalyzed for effects on additional secondary outcomes, such as happiness or sleep quality, at a later time or by another group of researchers. Quasi-experiments, based on exogenous (often historical) factors that can be harnessed using methods developed

Table 1. Approaches for Improving Inferences Based on (Secondary) Data Analysis

Method	Description
Data-blind analysis	To avoid their data analyses being affected by preconceptions, particle physicists and cosmologists use blind analysis (MacCoun & Perlmutter, 2015): Aspects of the data are altered (e.g., random noise is added to data points, variable labels are shuffled), all analytic decisions are made on this altered data set, and finally the analysis is run on the real, original data. Such an approach could also be used by psychologists analyzing secondary (and also primary) data.
Cross-validation	In the context of machine learning, cross-validation is a standard approach to avoid the statistical model being overfitted to the data at hand. The data set is repeatedly split into training and test subsets; the training data are used to estimate the model parameters, whereas the test data are used to evaluate the performance of the model (see Yarkoni & Westfall, 2017, for an introduction). If there are additional analytic flexibilities in model specification (e.g., decisions about which variables to include), this method can be expanded to nested resampling (Varma & Simon, 2006), in which analytic decisions are based on a separate part of the data, and the model is then estimated and evaluated (using cross-validation) on the remaining part of the data.
Holdout data	The very nature of secondary data opens the door to one highly effective mechanism to avoid overfitting: Data curators could hold back parts of the data. Researchers could then use the data available to them to specify and estimate their models, and the holdout data, provided after the completion of this initial analysis, could be used to obtain an unbiased estimate of model performance (suggested by Arslan, 2017). For example, in the Fragile Families Challenge (2017), researchers received access to parts of a longitudinal data set with more than 10,000 variables and were challenged to predict parts of the data they had not seen. To our knowledge, no major data holder or curator has yet implemented systematic holdouts, but this might be a promising future avenue.
Adjusted alpha level	Another approach to limit false-positive findings is setting a conservative alpha level. For example, researchers might want to use a level of .005 instead of .05 (Benjamin et al., 2018), or decrease their alpha as a function of sample size to balance error rates (Lakens, 2018). Note that this suggestion is by no means limited to secondary data analysis.
Coordinated analysis	The existence of multiple, independent, large-scale survey studies also allows for evaluation of generalizability in the context of secondary data analyses. In this kind of multicohort coordinated analysis (suggested by Hofer & Piccinin, 2009), researchers can test the same (or similar) analytic models in different samples, representing, for example, different geographic locations or cohorts, or different measurement instruments. Results can be pooled to better estimate an effect size and evaluate heterogeneity across differences in populations and methods.
Exploratory data analysis	All the preceding recommendations are applicable to confirmatory data analysis, but it is also important to consider exploratory methods. It has been argued that a major flaw of the way research is currently reported is that exploratory research is often written up as if it were confirmatory all along (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Clearly identifying exploratory analyses helps readers better assess the robustness of a particular result and opens the door for high-quality confirmatory follow-up research. We recommend that researchers omit *p* values and other tests of significance from exploratory analyses, as these cannot be interpreted properly without a confirmatory framework.

in econometrics, also bring an experimental aspect to secondary data analysis. For example, the Lanham Act of 1940 provided free, universal child care in the United States during World War II. Using U.S. Census data, researchers were able to track cohorts' outcomes to estimate the effect of this policy and found that it was associated with a strong and persistent increase in well-being (Herbst, 2017). Methods such as use of instrumental variables and regression discontinuity analysis can, when their assumptions are met, allow causal inferences from correlational data (Kim & Steiner, 2016), and genetic versions of these techniques, such as Mendelian randomization, are bringing a new causal aspect to secondary-data studies in biomedicine and beyond (Pingault et al., 2018).

Advantages and Disadvantages of Secondary Data Analysis

Preexisting data, if appropriately analyzed, offer great advantages: They can help situate effects in real-world behavior and outcomes and in diverse samples—or at least samples more diverse than undergraduate psychology students (Machery, 2010)—thus offering increased generalizability. Such data can, in the case of meta-analysis, be used to refine estimates obtained in prior work. They can be used to investigate hard-to-detect effects thanks to sample sizes that often exceed what is feasible for laboratory studies and therefore allow high-powered statistical tests. They often enable cross-country and cross-cultural research of considerable scope.

Panel studies repeatedly assess participants over years, even decades, which allows for complex longitudinal modeling. Many panel studies are conducted by teams representing a variety of disciplines, including psychology, economics, epidemiology, sociology, and demography; often the resulting data sets contain unique combinations of explanatory and criterion variables. Researchers sampling from these data sets have the opportunity to pair constructs from disparate fields to generate novel research questions. In addition, psychology researchers can benefit from the influence of these other fields. For example, demographers may work to ensure sampling of various geographic locations or subpopulations, which allows for more accurate representations of a country (Henrich, Heine, & Norenzayan, 2010). Panel studies often receive the funding necessary to assess biomarkers of health, thereby giving researchers data, such as brain MRI measures or data on genetic variants, that can be used to study small-sized yet potentially meaningful relations between psychosocial and biological variables. The largest recent example of a biomedical panel study is UK Biobank (Collins, 2012), and researchers have used its data to make important progress in current understanding of, for example, genetic links to behavioral traits such as neuroticism (e.g., Luciano et al., 2018).

By including many variables in a single data set, researchers have space for creativity and for exploring a range of novel research questions. Collaboration with other lab members, or other labs, is an excellent learning opportunity for early-career researchers, as they navigate different interests, limited resources, and new technologies. The resulting data set, if shared, creates a resource that can be returned to again and again for exploration, teaching exercises, and new collaborations. Even those data sets originally collected for a single study can serve as teaching tools, opportunities to explore an idea, and prototypes for designing new studies. Of course, at all times, researchers using preexisting data ought to (a) take care to be transparent about prior knowledge of the data and previous analyses of the data and (b) take measures to ensure robust inference, as detailed in a later section.

In addition to the advantages of its potential uses, secondary data analysis is an efficient way to conduct research: Preexisting data are often free, or at least entail marginal costs compared with paying an equivalent number participants for their time. Researchers also do not need to allocate time or space to collecting data when they use preexisting data. This makes preexisting data an especially attractive option for researchers with limited resources, such as graduate students, postdoctoral fellows, researchers at teaching-oriented universities, and mentors of undergraduates writing theses. Indeed, from the perspective of science as an endeavor constrained by limited resources, *not* using preexisting data when they are available and suitable to answer a research question could be considered inefficient and wasteful.

However, secondary data analysis is not without disadvantages. When applying secondary data analysis to data collected by someone else, a researcher relinquishes control over many important aspects of a study, including the specific research questions that can be answered. It may appear obvious, but if the researcher is interested in the relations between A and B, then both A and B must be measured, with a certain degree of reliability and external validity. Unfortunately, particularly in the context of large-scale survey studies, these criteria may not always be met. Because of the breadth of such studies, the data collectors may opt for short, coarse, and potentially unreliable measures in order to save time. For example, despite the impressive size of the UK Biobank study, some of the cognitive tests included in the initial sweep, likely because they were bespoke tests with very short durations, had very poor reliability (Lyall et al., 2016). These issues may be lessened when researchers analyze their own data, which is frequently the case in secondary data analysis; however, researchers will still grapple with data that were designed to answer a question different from the specific one they are currently studying or that were not designed with any specific questions in mind. Certain constructs might not have been assessed, or the ordering of steps in the experimental procedure might prohibit the correct temporal analysis. Researchers interested in longitudinal work might also find that the infrequency of measurement occasions or the length of time between them does not fit their research question. Furthermore, conclusions are necessarily restricted to the populations included in the original study. In short, researchers must weigh the convenience and power of preexisting data against the limitations they impose on the analysis and research question. As is the case with any other research tool, secondary data analysis is best used in conjunction with other methods (see Munafò & Davey Smith, 2018, for discussion of “triangulation” of research findings across multiple lines of evidence).

Despite its potential, secondary data analysis has been eschewed by some researchers who argue that it leads to “research parasites”—researchers who do not produce new data but simply live off the data collected by others (cf. Longo & Drazen, 2016). This concern appears to be symptomatic of misaligned incentives in psychology: Researchers are not rewarded for collecting high-quality data, although such an incentive could defuse concerns that others will "cash in" on one's data-collection labor; instead, researchers are rewarded for presenting striking results. Whereas many reforms are currently aimed at incentivizing better analyses and transparency (e.g., badges for open practices; Blohowiak et al., 2019; see the next section), the psychology community should consider building incentives for researchers who collect high-quality data and share it with others. For example, a data set archived in a public repository such as Dataverse could be equivalent to a publication on a curriculum vitae; and if other researchers use the data set in a productive manner, this downstream impact should be credited. The evaluation of job or tenure candidates could include attention to indicators of the quality of their data-collection efforts, such as the quality of measurement or the use of repeated measures or large samples. Fully acknowledging the collection of high-quality data as an integral contribution to science might require further development of data-sharing norms; publicly available, high-quality data are of limited use without documentation that enables other researchers to use the data (see Scott & Kline, 2019, for a discussion).

Secondary Data Analysis Through the Lens of Open Science

The field of psychology broadly has entered a phase of reflection and reform, mainly motivated by an inability to replicate and reproduce many key findings (Pashler & Wagenmakers, 2012). Large-scale collaborative efforts to evaluate the replicability of psychological effects have focused almost exclusively on studies that used primary data collection and experimental methods (e.g., Nosek et al., 2015). This is to be expected; replications of such studies are easier to carry out because they typically involve smaller sample sizes and more controlled environments than, for example, longitudinal cohort research. Researchers replicating lab-based research can more easily achieve high power and directly copy the testing conditions in the original experiment. We applaud these efforts, which have shed a great deal of light on which psychological findings can be relied upon and under which circumstances. But a consequence of the focus on experimental studies is uncertainty regarding the replicability of secondary research.

The replicability of an effect cannot be assessed until one is sure that the effect is *reproducible*. Whereas *replicability* refers to the extent to which a researcher can find the same effect with different data, reproducibility is the extent to which a researcher can find the same effect with the same data. Reproducibility is a key feature of transparent and robust research, as it results from well-documented analyses. To our knowledge, no one has tried explicitly to estimate the reproducibility of psychological effects found through secondary data analysis. However, such an attempt has been made in the field of economics, where secondary data analysis is the norm. Chang and Li (2018) found that of more than published 60 studies, fewer than half were reproducible, and assistance from the original authors was required to achieve reproducibility in many of these cases. Economics journals typically require the submission of code along with a manuscript, a practice that has not yet become mainstream in psychology. This leads us to predict that the reproducibility of psychological findings based on secondary data analysis will be lower than that in economics research.

As a necessary (but not sufficient) step to address issues of reproducibility and replicability, many scientists have advocated for the broad adoption of open-science values and practices (e.g., Klein et al., 2014; Nosek et al., 2015), most often implemented through disclosure and transparency in various forms. For example, one of the practical reforms of open science is the implementation of badges. These visual icons are attached to a published article along with links to online resources to signal that open-science practices have been used in the reported studies. The current set of badges—for open data, open materials, and preregistration (Kidwell et al., 2016)—have been adopted by a number of psychology journals, including *Psychological Science* (Eich, 2014) and *Advances in Methods and Practices in Psychological Science*. More generally, psychologists have outlined practices for all members of the scientific community, including researchers, teachers, and journal editors, to adopt in service of increasing the quality of research (Asendorpf et al., 2013; Funder et al., 2014; Lakens & Evers, 2014; van Assen, van Aert, Nuijten, & Wicherts, 2014).

Whereas the adoption of open-science practices appears to have increased the transparency of psychological science generally (Kidwell et al., 2016), the focus on laboratory-based methodologies has largely neglected the challenges faced by researchers using preexisting data. For example, if preexisting data are used, most—if not all—journal badges are unattainable (or introduce new ethical complications; Finkel, Eastwick, & Reis, 2015). The Open Data badge is often unavailable because access to the data from most panel studies requires registering with study coordinators, and data-sharing agreements prohibit sharing data among unregistered researchers. The Open Materials badge often cannot be awarded because many studies, especially those initiated decades ago, make use of copyrighted measures that are not permitted to be shared online. Finally, the Preregistration badge hinges upon posting analytic plans before data collection. Even if researchers do not analyze the data prior to registering an analytic plan, they cannot definitively prove—for example, with time-stamped variables—that they have not "peeked" at the data (run a few indicative tests) before making their hypotheses, nor can they prove that they have not read other studies that used the data to address similar questions. For these reasons, some people believe that secondary data analysis cannot be preregistered, although, as we make the case later in this article, this need not be true.

The relative difficulty of earning these badges when studies are based on secondary data analysis is in part a limitation of secondary-analysis projects: The trade-off of skipping the data-collection step is a lack of control over the component materials of the data used and, sometimes, a violation of traditional statistical assumptions. Yet it is these very concerns that have largely been ignored in the early discussions of open science and the development of methods and incentives for improving the transparency and robustness of psychological research. Few tools have been developed for the transparent and robust analysis of secondary data—a situation that falsely gives the impression that this type of research cannot be improved.

The implicit (and sometimes explicit) exclusion of secondary data analysis from open-science practices is unfortunate: As do all scientific endeavors, secondary data analysis in practice comes with many pitfalls and could be further improved if these were addressed. Aside from issues such as the lack of experimental control and the resulting restrictions on causal interpretation, secondary data analysis comes with a number of problems familiar to followers of the "replicability crisis" (Pashler & Harris, 2012). For instance, given the proliferation of variables in many of these data sets, it is all too easy to *p*-hack one's way to statistically significant, eye-catching results (Simmons, Nelson, & Simonsohn, 2011). This can be done in a variety of ways: For example, outcome switching, a practice common in clinical trials (Chan & Altman, 2005), is also prevalent (in our experience) in secondary analysis; tests of interactions between potential predictors can be added on a whim; and researchers can simply plug in one covariate after another until a significant result or the desired effect size is obtained.

Another common problematic practice is subgroup analyses. Sometimes this practice is obvious—for example, when specific ethnic groups are examined separately—but subgroup analyses can be less conspicuous. For example, researchers may choose to analyze data from a single wave of a longitudinal panel study, deliberately or otherwise ignoring variables collected at another wave that have important statistical or theoretical links to the constructs of interest.

In the case of repeated measures, researchers can examine multiple cross-sectional relationships and present only the significant results. Certainly these kinds of practices are possible in most studies (Simmons et al., 2011), but in analyses involving large, preexisting data sets, the temptation to "try it" with another variable or subgroup—selected post hoc—is often strong, and the large sample sizes involved mean that perseverance is likely to be "rewarded" with a *p* value below the alpha level for significance or a substantially large effect size. As a result, researchers using large data sets are more likely than those using small data sets to present models that fit random variation in their data—especially as models increase in complexity—instead of revealing reliable, generalizable associations. That is, they are more likely to overfit their models to the data and reduce the potential for replication of their results.

Unique to secondary data analysis is the problem of familiarity with the data. A key reason for using a preexisting data set is that it may be the sole source of data appropriate for evaluating a particular research question. For example, questions about life-span development require decades of data, such as the unique life-span data from the Lothian Birth Cohorts (Deary, Gow, Pattie, & Starr, 2012). Biomarker and genetic data often require a very large team of research assistants, medical professionals, and data scientists (found in large quantities in few studies other than UK Biobank; Collins, 2012). Given the limited number of data sets available to answer questions in these areas, along with the huge number of variables available in existing data sets, it is expected that researchers will return to the same data sets multiple times to investigate different (but similar) research questions. Unfortunately, this practice introduces biases, because researchers become aware of relations in the data. Consequently, they can design complex models that fit the data with very few changes or propose very specific hypotheses that are substantiated with few caveats. These are not truly predictions, because the researchers already had some knowledge of how the variables related to one another. As pointed out by Gelman and Loken (2013), the problem is not necessarily the number of ways researchers analyze their data, but rather the number of potential ways they *could* do so. When researchers make analytic decisions based on their data rather than their theory, the multiple potential comparisons must be taken into account when interpreting the results.

The proliferation of published research using these data sets means that even a researcher who has never worked with a particular data set before will likely have some knowledge of the patterns within it. Many preexisting data sets have been repeatedly mined in this way, usually by scholars in the same subfield. For example, the Health and Retirement Study (HRS; Juster & Suzman, 1995) has been used by personality psychologists studying smoking (Weston & Jackson, 2015), longevity (Hill, Turiano, Hurd, Mroczek, & Roberts, 2011), and biomarkers of health (Luchetti, Barkley, Stephan, Terracciano, & Sutin, 2014). It is to be expected that these researchers will read each other's scholarly work, because it provides substantial information for generating and testing theories concerning relations between health and psychology. However, in the process of developing well-grounded hypotheses, these researchers also become aware of relationships in the HRS data set, regardless of whether they have previously analyzed these data, and are potentially biased by what they have learned. This curse of knowledge does not preclude researchers from analyzing a data set they have read about. But such prior knowledge can bias researchers' choices regarding which research questions to ask, how to wrangle variables, and how to fit models.

Because of the opportunity to capitalize on researcher degrees of freedom and the increased likelihood of results-biased decision making, research employing secondary data analysis must be held to as high a standard as research using primary data collection—if not a higher one. In what follows, we make recommendations for (a) increasing the robustness of secondary data analysis by increasing its transparency and (b) estimating the robustness of the results obtained. We make these suggestions to researchers who value open science and wish to produce research that will stand the tests of time and replication. However, it is our hope that these recommendations will inspire journal editors, grant reviewers, tenure committees, and everyone who has the formal power to change incentives in the scientific community.

Recommendations for Transparent Secondary Data Analysis

Increasing transparency is a cornerstone of the current open-science reform movement. The objective of attempts to increase transparency is to live up to the ideal summarized in the motto of the United Kingdom's Royal Society: *Nullius in verba*, or "take nobody's word for it." Scientists need not be taken at their word when all their materials, methods, and actions are available for anyone to see. The badges we have described all traffic in transparency: Data and materials are the ingredients of a study, and preregistration clarifies which analytic decisions were made before the authors knew anything about the data (or results from the data) and which were not. This last point is key. If data-analytic decisions are based on the collected data themselves, then traditionally used statistical tests can no longer successfully control error rates.

The tendency to make decisions based on data rather than theory becomes more likely, maybe even certain, in the case of preexisting data, especially if a researcher has used or even read about the data in the past. Take, for example, the proliferation of publications reporting analyses of the HRS data. During a thorough literature review, personality-and-health researchers will read frequently about this data set and become aware that the traits of extraversion and conscientiousness are highly correlated in the HRS. They may therefore choose to use conscientiousness as a covariate when examining the relationship of extraversion to health. This alone is not problematic; the problem is that when the study is published, readers will have no way to know that this decision was based on prior knowledge. Transparency clarifies for readers of science which decisions were theory based and independent of the data and which were not, and this allows them to interpret results appropriately. More specifically, readers (and the researchers conducting a study) should have less confidence in analytic results when analyses were designed, in part, on the basis of prior knowledge of the data than when analyses were designed without such prior knowledge. We recommend several ways in which researchers can transparently document a secondary data analysis:

First, researchers can provide links to codebooks and instructions for accessing the data. If the preexisting data set is from a panel study or available for purchase, there are likely to be publicly available codebooks or websites where the data can be accessed. These can helpfully supplement the Method sections of publications reporting analyses of the data set as well as workflows based on, for example, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (von Elm et al., 2007). If researchers own the data set, they can create their own codebook with relevant information (e.g., Vardigan, Heus, & Thomas, 2008; for an example, see Condon & Revelle, 2015). If the data are not their own, they can describe how they were able to access them. Panel studies and data banks often e-mail researchers when they have provided access to some or all of the data. Copies of this correspondence, or any data-access statements, can be made available to readers as supplementary material. Such correspondence often contains a date, which can be important time-stamp information if the researchers chose to register analyses prior to accessing the data (see our discussion of preregistration later in this section).

Second, researchers should communicate how the data have been used. This recommendation is not meant to prohibit researchers from using the data set again. The point is to reconstruct this context, which is next to impossible if it is not done incrementally. Enumerating prior experience with a data set openly simply makes clear both to the readers and to the researchers themselves how much prior knowledge went into generating hypotheses or designing models. Regarding work by other people, researchers might simply document the instances they have come across during their literature review. A thorough description of the prior literature is likely central to developing a research hypothesis and writing an article's introduction section, so we recommend integrating this description into a literature review prior to conducting the analyses, as this will likely save time when writing up the results. In the case of the researchers' own previous research, citations to past publications that are pertinent to the research question at hand should be provided.

A note of caution is warranted: It is quite likely that a researcher's history analyzing a particular preexisting data set is not limited to what has been published. Researchers should disclose any analysis that is relevant to the current project. Specifically, this includes the calculation of any statistic or the creation of any visualization that includes at least one variable in the project. We believe that this process will become easier as preregistration and preprints are more widely and consistently used. Ideally, it will become relatively easy to use a platform such as the Open Science Framework to link to prior projects that carefully document both published and unpublished analyses of a preexisting data set. Today, however, this is not an easy task for most researchers. Because preregistration and preprints have only recently been adopted in psychological science, this task may actually prove impossible in some cases. There are no easy solutions for ensuring and checking that researchers have disclosed all knowledge of a data set. Unfortunately, this creates opportunity for motivated naïveté and strategic laziness. We must therefore acknowledge that this recommendation—disclosing all prior knowledge—addresses only part of the problem. We hope that work continues on this front.

Third, researchers can document the data-wrangling and -analysis pipeline. Sharing the analytic script is not always considered part of sharing materials, depending on the journal, but it is especially important for researchers using preexisting data. A key component of secondary research is documenting by way of code and precise instruction the steps required to access, merge, and prepare the data prior to formal statistical analysis. These procedures are often extremely complex. Moreover, important details are often left out of academic publications and are dependent on the time of data access and the exact version of the data that was accessed. As models increase in their complexity, it often becomes more difficult to describe to readers how data were modified and analyzed, especially given the space constraints of many journals' Method sections (e.g., as Chang & Li, 2018, found). Sharing the analysis script instantly deals with this problem.

Fourth, we recommend that secondary data analysis be preregistered. As in the case of primary data analysis, preregistration should occur before the analyses are conducted. Preregistration forms should enumerate any planned analyses and all analytic decisions related to those analyses, for example, the numeric definition of outliers and the procedure for handling them, or how a particular measure will be scored. Researchers can also preregister analyses for upcoming waves of publically available data sets. We note that this system could be expanded to exploratory data analysis as well: The preregistration could simply note a plan to explore relationships between specified variables. At the time of this writing, the Center for Open Science is developing an interactive form for preregistering analyses using preexisting data. There is also a template Open Science Framework project (Weston et al., 2018) that guides researchers through the information relevant for preregistering secondary data analyses.[3]

Applying the term *preregistration* to the analysis of preexisting, potentially accessible, secondary data is somewhat controversial. Some researchers have argued that the term should be reserved for registration of studies prior to data collection (e.g., Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015). One of the arguments is that there is no way to prove definitively that a researcher has not looked at the data (or results from the data) prior to analysis.

Preregistrations are very much an imperfect business at present. Many preregistration protocols are too vague to safeguard against *p*-hacking (Wicherts et al., 2016); the published manuscript might not follow the prespecified analysis plan, nor is it clear how or whether journals should evaluate fidelity to an analysis plan (Tucker, 2014). Reviewers and readers must still compare the preregistration with the final study to evaluate adherence, and adherence itself tells little about other important aspects of a study's quality (e.g., the validity of the design). Hence, one could argue that a preregistration per se does not imply much, which is why the label should not be interpreted as a signal of superior quality. We note that preregistration is not a box-ticking exercise. In evaluating a manuscript, attention should be paid not just to whether a preregistration exists, but also to the content and quality of that preregistration.

Given that preregistration seems to have acquired a special definition referring to registration prior to data collection, and given its prominent role in the "Open Science Trifecta" (Open Data, Open Materials, Preregistration), researchers who rely on secondary data may assume incorrectly that open science is irrelevant or inaccessible to them. A simplistic “preregistration or it didn’t happen” mind-set might even lead researchers to conclude that secondary data analysis is second-class research because it cannot be fully preregistered, and thus might widen the chasm between different research traditions.

Hence, we argue that the term *preregistration* can in fact be applied to secondary data analysis, mostly for pragmatic reasons, and at the same time, we would like to encourage more discussion about what preregistrations can and cannot achieve, in the context of both primary and secondary data analysis. For example, preregistrations are always trust based regardless of whether the data already existed, because it is possible to “pre”-register a study that has already been conducted, and because there is currently no mechanism in place that prevents researchers from filing multiple preregistrations (potentially on different platforms) with slightly different analysis plans and later selectively reporting the one that “worked.” Scientists who yearn for a bulletproof approach that cannot be gamed by insincere authors might prefer adapting the Registered Report format to studies with preexisting data, which (a) would make it very hard to “pre”-register secondary data analyses that have already been performed because reviewers’ feedback during the initial stage can lead to substantial changes in analyses and (b) partially remove the incentive to produce a certain result thanks to in-principle-acceptance prior to data collection and analysis. Registered Reports, unlike weaker preregistration of analysis plans, might preclude secondary data analysis when researchers cannot supply evidence that they had no prior access to the data, although this too should be a point of discussion.

Recommendations for Improved Inference Based on Secondary Data Analysis

Researchers often face a large number of decisions while analyzing their data (e.g., whether and how to transform variables, which covariates to include, which estimator to use), and they might often genuinely be unsure about the best statistical approach for their research question. This is even more of an issue with data sets that are rich in variables. Thus, in the context of secondary data analysis, the robustness of findings becomes a central concern: Would conclusions substantially change if a different plausible model specification were used?[4]

On the basis of empirical testing, Young and Holsteen (2017) described three different degrees of model robustness: First, the result may hold no matter how the model is specified (i.e., the finding is robust). Second, the result may depend on some specific model ingredients, such as a particular covariate (i.e., there is systematic variability). Third, the result may depend on a very specific combination of parameters and arise only in one (or a few) of many possible models (“knife edge” specification). A robust finding increases confidence that conclusions are not based on a fluke. Systematic variability calls for follow-up analyses to clarify the role of the critical model ingredients. Knife-edge specifications call for prudence: If only one in a multitude of plausible models supports a particular finding, that finding is likely a mere fluke in the data.

The simplest way to probe the robustness of a finding is to perform robustness checks (also known as sensitivity analyses), which are a staple in economics research but appear to be less common in psychology (see, e.g., Duncan, Engel, Claessens, & Dowsett, 2014, for a comparison of journals in economics and those in developmental psychology).[5] In the most standard kind of robustness check, the model is rerun with one element of the specification changed. Reports of robustness checks might range from a simple footnote (e.g., “results remained unaffected when age was included as a covariate”) to a supplementary website presenting results from dozens of models in such a way that they can be compared by the reader (e.g., Arslan et al., 2017).

The fundamental idea of robustness checks can be expanded to include checking all possible combinations of all plausible model ingredients. Naturally, this quickly leads to rapid growth of the number of possible models: For example, if there are only three simple dichotomous decisions to be made (control for gender or not, control for age or not, remove outliers or not), 8 different model specifications (2 × 2 × 2) result. Several researchers have advocated that all these models should be run and reported. For example, Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016) labeled this approach a *multiverse analysis*; Young (2018) described it as the *computational solution* to model uncertainty; and Simonsohn, Simmons, and Nelson (2015) developed the concept of *specification-curves* analysis, which allows researchers to calculate a p value across all specifications.

Specification-curve analysis has been successfully applied in at least two investigations built on large-scale social data sets: a study of birth-order effects on personality (Rohrer, Egloff, & Schmukle, 2017) and a study on the impact of digital-technology use (gaming and use of social media) on psychological well-being (Orben & Przybylski, 2019a). Results derived from the work on birth-order effects suggested that some published results in this area might depend on knife-edge specifications. The thorough analytic approach to the effects of gaming and social media (a) provided a robust estimate of their modest impact on young people and (b) used the inherent richness of the available data to put the effect sizes within clear everyday contexts (e.g., by comparing the associations between technology use and well-being with the associations between potato consumption and well-being). Such context is absolutely necessary when comparing two analyses of the same preexisting data set that arrive at divergent conclusions. Specification-curve analysis can reveal that researchers have poked around a large-scale social data set in a nonsystematic way. Eye-catching correlations are easily publishable, and specification-curve analysis can reveal cases in which researchers have selected extreme pairings of predictors and outcomes (for an illustrative example, see Supplementary Table 6 from Orben & Przybylski, 2019b).

Though we recommend the use of robustness checks and their expansions, they are still no guarantee that the data have not been overfitted. Among economists, one sometimes hears jokes about how wondrous it is that robustness checks always work to confirm the finding; the danger of selectively including only model ingredients that support one's preferred conclusion is certainly higher than zero. Hence, to further strengthen robustness checks, we recommend that they also be preregistered.

Beyond robustness checks, there are additional approaches that can be used to ensure that biases do not affect secondary data analyses and to avoid overfitting. We have included some of these approaches in Table 1. We note that these recommendations are not specifically for secondary data analysis and are used with great success in analyzing primary data.

Into the Future

We have recommended methods to ensure that secondary data analysis is transparent and that reported results from secondary data analysis are robust. We finish with three calls to action.

First, we encourage researchers who run laboratory experiments with the potential for further analyses to consider making their data sets available for other researchers to analyze as well. Such data sets—whether made completely open or accessed with permission—constitute valuable resources for future research. We believe that the production and curation of such data sets should be considered a research output with value akin to publications or developing statistical software packages.

Second, we turn to subfields of psychology in which secondary data analysis is frequently used, such as personality, individual differences, and developmental psychology. The use of secondary data analysis—specifically, the use of a few large surveys—can create the illusion of replication or convergence across an area of research. We say "illusion" because a large proportion of published results within a field may be based on the same panel study or data set. This could result in a growing literature in which a relatively large number of publications report similar effects, but the number of truly independent tests does not expand. For example, the German Socio-Economic Panel Study has repeatedly been used to track personality development, including twice in the same issue of the same journal (Lucas & Donnellan, 2011; Specht, Egloff, & Schmukle, 2011). This is not necessarily problematic—and could even be beneficial to probe the robustness of different analytic approaches—if it is clear to readers that the same data are being used to answer similar or sometimes identical questions. However, without better indexing for data (e.g., clear tags referring to the data source), the extent of use of a particular data set is difficult for readers to evaluate.

For example, how much of the evidence for the link between trait conscientiousness and health is based on data from the HRS? Dependence between published findings limits certainty in an effect. If a published literature is largely supported by one data set, or even a small number of data sets, rather than by a large number of data sets, one should be less certain that the effect in question is generalizable to other samples. Multiple related findings from a single data set may not suggest multiple independent effects, but rather may reflect one effect with shared variance across a number of indicators. For example, conscientiousness has been found to be associated with mortality (Hill et al., 2011), the incidence of chronic conditions (Weston, Hill, & Jackson, 2015), health behaviors (Hakulinen et al., 2015; Roberts, Smith, Jackson, & Edmonds, 2009), and sleep (Hintsanen et al., 2014), but each of these studies used the HRS data set. Consequently, each of these publications cannot be counted as reporting an independent result: They all relied on the same sample of individuals as well as on the same operationalization of conscientiousness. Without independent verification of each of these associations in new data sets, this evidence merely suggests that conscientiousness, as measured in the HRS, may be related to some (likely overlapping) aspects of health in the HRS sample. We call for introspection and systematic review of key findings in the subfields of psychology, especially for those findings about which psychological scientists feel certainty. Are these effects found in multiple, independent data sets? Or are they found in only one or two data sets, over and over again?

We also call for systematic reproducibility checks on studies using secondary data analysis. How many researchers have preregistered these studies, or made code available, or in any way ensured that other researchers can readily reproduce the effects? We especially appeal to academic journals, which could hire statistical editors or reviewers whose jobs are to reproduce analyses and results using the code and data provided or specified. At the time of this writing, this step has already been taken by six academic journals, including *Meta-Psychology*, and so a viable model exists (for an example, see the Data Reproducibility Policies section of Mellor, Esposito, DeHaven, & Stodden's, 2016, wiki page at osf.io/kgnva).

Third, and finally, we turn to readers who are interested in developing technologies for the advancement of open science and call for the development of tools specific to secondary data analysis. Certainly there are ways to adapt existing tools for such work (e.g., preregistering secondary data analysis). But in the case of preexisting data, there are specific challenges that can be addressed. Other researchers have proposed data-checkout systems as a form of preregistration and monitoring of data use (Scott & Kline, 2019). We also suggest the development of tools for tracking and reporting prior knowledge of a data set.

Conclusion

Our purpose here is to urge psychologists to consider secondary data analysis as a powerful and low-cost tool for exploring important research questions. We suggest that researchers with limited resources, especially, should consider the ways in which preexisting data might supplement or form the foundation of research and teaching programs. Secondary data sets have many strengths, but researchers capitalizing on their value can easily fall prey to several of the limitations and questionable research practices that still haunt psychological science amidst the replication crisis. Fortunately, many of the specific reforms that have begun to improve the credibility of primary research either can be directly implemented in secondary analysis or have analogues that can be used in secondary analysis. Preregistration can be implemented in secondary data analysis by registering analyses before data are accessed or before results from analyses are known. Inferences can be improved using strategies such as cross-validation, holdout samples, and multicohort analyses. The robustness of results can be thoroughly tested in a multiverse analysis. We present this article as a manifesto for the improvement of secondary data analysis, to ensure that this critically important type of research is carried along with the open-science revolution.

Action Editor

Simine Vazire served as action editor for this article.

Author Contributions

All the authors contributed to writing and editing this manuscript.

ORCID iDs

Sara J. Weston https://orcid.org/0000-0001-7782-6239
Julia M. Rohrer https://orcid.org/0000-0001-8564-4523
Andrew K. Przybylski https://orcid.org/0000-0001-5547-2185

Acknowledgments

None of this work would have been possible without the contributions of the participants in a discussion of open-science strategies for secondary data analysis that took place at the second annual meeting of the Society for the Improvement of Psychological Science, in Charlottesville, Virginia, in August 2017. Those participants were Cassandra Brandes, William Chopik, Patrick Forscher, Nicholas Fox, Lisa Hoplock, Rich Lucas, Kathryn Mills, John Protzko, Kathleen Reardon, Kate Rogers, and Kaitlyn Werner.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

A. K. Przybylski was supported by an Understanding Society Fellowship funded by the Economic and Social Research Council (ES/K005146/1). The funders had no role in the decision to publish or preparation of the manuscript.

Open Practices

Open Data: not applicable
Open Materials: not applicable
Preregistration: not applicable

Notes

[1] We recommend that readers browse the "Cohort Profile" section in each issue of the *International Journal of Epidemiology* for details on a huge number of other such data sets.

[2] The term *secondary data* is sometimes used to refer to data that are collected by one researcher (or team of researchers) and analyzed by a second researcher (or team; e.g., Vartanian, 2010). We choose not to use this definition because preexisting data may have been collected by the same researchers who wish to analyze those data. However, we retain the use of the term *secondary data analysis* to connect our work with that of other researchers who have sought to improve the robustness of research using secondary data and have curated lists of available data sets.

[3] In addition, at a meeting of the Society for the Improvement of Psychological Science in July 2018 (Grand Rapids, MI), a group of researchers built on the principles and recommendations of this manuscript and developed a template for registering secondary data analyses. Those researchers were Olmo van

den Akker, Marjan Bakkar, Brian Brown, Lorne Campbell, William Chopik, Oliver Clark, Rodica Damien, Pamela Davis-Kean, Charlie Ebersole, Andrew Hall, Matthew Kay, Jessica Kosie, Elliot Kruse, Jerome Olsen, Stuart Ritchie, Courtney Soderberg, K. D. Valentine, Anna Van't Veer, and Sara J. Weston.

4. Of course, robustness can also be a central concern in primary data analysis, as illustrated in Credé and Phillips (2017).

5. It should be noted that these authors used a slightly different definition of *robustness check* that included the replication of a finding using a new data set.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*, 235–244.

Arslan, R. C. (2017, September 14). Overfitting vs. open data [Blog post]. Retrieved from http://www.the100.ci/2017/09/14/overfitting-vs-open-data/

Arslan, R. C., Willführ, K. P., Frans, E., Verweij, K. J. H., Bürkner, P., Myrskylä, M., . . . Penke, L. (2017). *Paternal age and offspring fitness: Online supplementary website* (Version v2.0.1). doi:10.5281/zenodo.838961

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Blohowiak, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., Hasselman, F., . . . DeHaven, A. C. (2019). *Badges to acknowledge open practices*. Retrieved from osf.io/tvyxz

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). *Registered Reports*: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Chan, A.-W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *British Medical Journal*, *330*, 753. doi:10.1136/bmj.38356.424606.8F

Chang, A. C., & Li, P. (2018). Is economics research replicable? Sixty published papers from thirteen journals say "often not." *Critical Finance Review*. Advance online publication. doi:10.1561/104.00000053

Collins, R. (2012). What makes UK Biobank special? *The Lancet*, *379*, 1173–1174.

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3*, Article e6. doi:10.5334/jopd.al

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological & Personality Science*, *8*, 493–499.

Deary, I. J., Gow, A. J., Pattie, A., & Starr, J. M. (2012). Cohort profile: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*, *41*, 1576–1584.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*, 2417–2425.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. doi:10.1177/0956797613512465

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275–297.

Fragile Families Challenge. (2017). Retrieved from http://www.fragilefamilieschallenge.org/

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Hakulinen, C., Hintsanen, M., Munafò, M. R., Virtanen, M., Kivimäki, M., Batty, G. D., & Jokela, M. (2015). Personality and smoking: Individual-participant meta-analysis of nine cohort studies. *Addiction*, *110*, 1844–1852.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral & Brain Sciences*, *33*, 111–135.

Herbst, C. M. (2017). Universal child care, maternal employment, and children's long-run outcomes: Evidence from the US Lanham Act of 1940. *Journal of Labor Economics*, *35*, 519–564.

Hill, P. L., Turiano, N. A., Hurd, M. D., Mroczek, D. K., & Roberts, B. W. (2011). Conscientiousness and longevity: An examination of possible mediators. *Health Psychology*, *30*, 536–541.

Hintsanen, M., Puttonen, S., Smith, K., Törnroos, M., Jokela, M., Pulkki-Råback, L., . . . Keltikangas-Järvinen, L. (2014). Five-factor personality traits and sleep: Evidence from two population-based cohort studies. *Health Psychology*, *33*, 1214–1223.

Hofer, S. M., & Piccinin, A. M. (2009.). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*, 150–164. doi:10.1037/a0015566

Juster, F. T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, *30*, S7–S56.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS*

Biology, *14*(5), Article e1002456. doi:10.1371/journal.pbio.1002456

Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychology*, *51*, 395–405. doi:10.1080/00461520.2016.1207177

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, USA*, *110*, 5802–5805.

Lakens, D. (2018, December 1). Justify your alpha by decreasing alpha levels as a function of the sample size [Blog post]. Retrieved from http://daniellakens.blogspot.com/2018/12/testing-whether-observed-data-should.html

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278–292.

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *New England Journal of Medicine*, *374*, 276–277. doi:10.1056/NEJMe1516564

Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: Longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology*, *101*, 847–861.

Luchetti, M., Barkley, J. M., Stephan, Y., Terracciano, A., & Sutin, A. R. (2014). Five-factor model personality traits and inflammatory markers: New data and a meta-analysis. *Psychoneuroendocrinology*, *50*, 181–193.

Luciano, M., Hagenaars, S. P., Davies, G., Hill, W. D., Clarke, T.-K., Shirali, M., . . . Deary, I. J. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics*, *50*, 6–11.

Lyall, D. M., Cullen, B., Allerhand, M., Smith, D. J., Mackay, D., Evans, J., . . . Pell, J. P. (2016). Cognitive test scores in UK Biobank: Data reduction in 480,416 participants and longitudinal stability in 20,346 participants. *PLOS ONE*, *11*(4), Article e0154222. doi:10.1371/journal.pone.0154222

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*, 187–189.

Machery, E. (2010). Explaining why experimental behavior varies across cultures: A missing step in "The weirdest people in the world?" *Behavioral & Brain Sciences*, *33*, 101–102.

Mellor, D., Esposito, J., DeHaven, A., & Stodden, V. (2016). *Transparency and Openness Promotion (TOP) guidelines*. Retrieved from https://osf.io/kgnva/

Munafò, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature*, *553*, 399–401.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 21. doi:10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. doi:10.1126/science.aab2374

Orben, A., & Przybylski, A. K. (2019a). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, 173–182.

Orben, A., & Przybylski, A. K. (2019b). The association between adolescent well-being and digital technology use [Supplemental material]. *Human Behaviour*, 3. Retrieved from https://static-content.springer.com/esm/art%3A10.1038%2Fs41562-018-0506-1/MediaObjects/41562_2018_506_MOESM1_ESM.pdf

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Pingault, J. B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., & Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, *19*, 566–580. doi:10.1038/s41576-018-0020-3

Roberts, B. W., Smith, J., Jackson, J. J., & Edmonds, G. (2009). Compensatory conscientiousness and health in older couples. *Psychological Science*, *20*, 553–559.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42.

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science*, *28*, 1821–1832.

Scott, K. M., & Kline, M. (2019). Enabling confirmatory secondary data analysis by logging data checkout. *Advances in Methods and Practices in Psychological Science*, *2*, 45–54.

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. Retrieved from http://dx.doi.org/10.2139/ssrn.2694998

Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, *101*, 862–882.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Tucker, J. (2014, September, 18). Experiments, preregistration, and journals [Blog post]. Retrieved from https://blog.oup.com/2014/09/pro-con-research-preregistration/

University of Essex, Institute for Social and Economic Research. (2018). British Household Panel Survey: Waves 1-18, 1991-2009 [Data set]. doi:10.5255/UKDA-SN-5151-2

U.S. Bureau of Labor Statistics. (n.d.). *National Longitudinal Surveys*. Retrieved from http://www.nlsinfo.org/

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLOS ONE*, *9*(1), Article e84896. doi:10.1371/journal.pone.0084896

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, *3*(1), 107–113.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, Article 91. doi:10.1186/1471-2105-7-91

Vartanian, T. P. (2010). *Secondary data analysis*. Oxford, England: Oxford University Press.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P., for the STROBE Initiative. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Annals of Internal Medicine*, *147*, 573–577.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). Enhancing the power of household panel studies: The case of the German Socio-Economic Panel Study (SOEP). *Schmollers Jahrbuch*, *127*, 139–169.

Weston, S. J., Hill, P. L., & Jackson, J. J. (2015). Personality traits predict the onset of disease. *Social Psychological & Personality Science*, *6*, 309–317.

Weston, S. J., & Jackson, J. J. (2015). Identification of the healthy neurotic: Personality traits predict smoking after disease onset. *Journal of Research in Personality*, *54*, 61–69.

Weston, S. J., Mellor, D. T., Bakker, M., Van den Akker, O., Campbell, L., Ritchie, S. J., . . . Nguyen, T. T. (2018). *Secondary data preregistration*. Retrieved from osf.io/x4gzt

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*, Article 1832. doi:10.3389/fpsyg.2016.01832

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, *4*. doi:10.1177/2378023117737206.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*, 3–40.