RUNNING HEAD: PROMOTING ROBUST BIG DATA RESEARCH

Promoting Robust and Reliable Big Data Research in Psychology

Joshua A. Strauss
James A. Grand

University of Maryland

Citation
Strauss, J.A., & Grand, J.A. (2020). Promoting robust and reliable big data research in psychology. In S.E. Woo, L. Tay, & R. Proctor (Eds.), *Big data in psychological research*. (pp. 373-392). Washington, DC: American Psychological Association.

Imagine being tasked with constructing the table of contents for a book whose goal was to summarize the trends, developments, and themes that defined psychological science over the past fifty years. When considering how to describe the past decade, one would be hard pressed to choose between the *rise of computational social science* (including Big Data research) and matters of *robust and reliable science* as the titular focus. In some respects, these topics have found themselves intertwined. For example, the accumulation and analysis of largescale replication datasets has been used as a tool for demonstrating concerns regarding the robustness of published findings in psychology (e.g., Camerer et al., 2018; Klein et al., 2014; Open Science Collaboration, 2015). However, increasing the volume of data used to examine psychological phenomena hardly registers on the scale of what excites most about the advent of Big Data and a more computationally-oriented social science. Interest in these new frontiers is encapsulated by the promise of new discoveries, insights, and the development of predictive tools for understanding human affect, behavior, and cognition that can be used to shape future knowledge generation and policy decisions.

Unfortunately, the excitement and potential of Big Data analytics and computational social science makes it all too easy to lose sight of the issues that have contributed to worries regarding the reliability and robustness of psychological research in general. Lazer, Kennedy, King, and Vespignani (2014) coined the term “Big Data hubris” to reflect the implicit belief that the use of large datasets and sophisticated analyses provides researchers license to relax principles of scientific rigor such as accurate measurement, construct validity, and reliability. Similar concerns have also been raised regarding replicability and reproducibility in Big Data research. For example, Leetaru (2017) recounts the many methodological challenges faced in attempting to replicate Big Data research, such as determining whether an original and

replication dataset are equivalent and whether or how the decisions/algorithms used by a researcher to collect, organize, and analyze Big Data can even be effectively reproduced. Still others have raised concerns regarding the transparency of Big Data research, in addition to more complex issues regarding societal and technical infrastructures (ever-changing government policies and business practices that influence data quality, programmers/developers dynamically restructuring data and data access protocols, etc.; Boyd & Crawford, 2012; Lazer et al., 2014).

These points should give pause to even the strongest advocates of Big Data analytics to consider how its unique strengths can be leveraged to advance psychological science and practice without repeating the sins of our past. The focus of the present chapter thus concerns a critical question—what can Big Data and computational social science do to improve the likelihood that its research meets emerging criteria for robust and reliable psychological science? In reflecting on this topic, we have elected not to debate the merits of specific methodologies and analyses available to Big Data researchers or when computational approaches may be more or less appropriate. These are clearly important matters; however, our intention is to discuss and provide guidance applicable to establishing norms and standardized practices for the conduct, reporting, and dissemination of Big Data research. We begin by first describing what we believe are the characteristics of a robust Big Data science and some of the more significant challenges for meeting these demands. The remainder of the chapter then focuses on three issues related to scientific credibility that have been frequent topics of discussion in psychology (hypothesizing after results are known (HARKing), questionable research practices (QRPs), and replicability/reproducibility, describing their relevance to Big Data research, and offering recommendations for facilitating reliable and robust contributions of Big Data science to psychology.

What Makes a Science Robust?

Although the replicability and reproducibility of findings in the social and psychological sciences has seemingly received the most widespread attention (Camerer et al., 2018; Fanelli, 2010a; Klein et al., 2014; Open Science Collaboration, 2015), virtually all disciplines of science wrestle with similar issues (e.g., Fanelli, 2009, 2011; Ioannidis, 2005a, 2005b; Marcus, 2014; Rubin, 2011). A great deal has already been written regarding the purported causes of the "credibility crisis" in science, including how both top-down/environmental forces (e.g., "publish or perish" norms and incentive structures in academia) and bottom-up/individual behaviors (e.g., engaging in research practices to "game" the system) across a variety of stakeholders in the scientific enterprise can collectively impact the trustworthiness of research (National Academies of Sciences, Engineering, & Medicine, 2017). Rather than reiterate those points again, we wish to adopt a more aspirational lens that elaborates what we believe constitutes a robust and reliable scientific field of inquiry and consider what that vision entails for research involving Big Data methods and analytics.

To frame this discussion, we rely on the defining characteristics of a robust science proposed by Grand, Rogelberg, Allen, et al. (2018). As opposed to a checklist or set of standards for judging individual researchers or pieces of scholarship, these characteristics are intended to distill the values that reflect "better science" (Grote, 2016) and serve as markers for evaluating how decisions, policies, resources, and/or practices intended to improve scientific credibility contribute to that goal. In the sections below, we define and apply these characteristics to research conducted using Big Data approaches. Table 1 provides a summary of this discussion.

Robust Big Data Science should be *relevant*

Grand, Rogelberg, Allen, et al. (2018) characterize relevance with respect to the utility of the research generated by a science. Specifically, a more robust science is one in which the knowledge produced by a discipline improves understanding of the natural world, can be used to address important needs, and builds towards contributions that benefit society. In many ways, the principle of relevance concerns the extent to which scientific outputs are problem-focused, solution-oriented, and attempt to "do good." Big Data applications should seemingly adhere to this principle well given that they are frequently described as tools for extracting evidence-based insights into complex and often intractable problem domains (e.g., Kim, Trimi, & Chung, 2014; Ryan & Herleman, 2015). However, when the generation of such insights occurs through inductive/exploratory methods (e.g., unsupervised learning techniques) and through the use of data sources/models not designed with an eye towards drawing the intended inferential claims or maintaining individual protections, the relevance and applicability of such knowledge should be appropriately vetted.

Lazer et al. (2014) provide an excellent commentary and case study on this challenge for Big Data research in the context of estimating the prevalence of flu cases using Google search activity. Developing a model capable of automatically and in near-real time predicting flu outbreaks is an admirable scientific pursuit with clear implications for positively influencing healthcare practice and policies. However, the Big Data model was frequently outperformed by and resulted in systematically biased overestimations compared to existing models that used and analyzed data using more "traditional" methods (e.g., local laboratory surveillance reports collected by the Centers for Disease Control and Prevention, simple time-lagged regression). Lazer et al. (2014) suggest this case study offers a number of important lessons into ensuring the relevance, utility, and trustworthiness of Big Data research and applications, including the need

to establish whether, how, and why the insights produced by these techniques improve upon existing knowledge. We echo this sentiment and the position that ensuring a robust Big Data science is relevant requires researchers explicate and monitor the purpose of their investigations (e.g., confirmatory vs. exploratory) and make concerted efforts to verify the veracity of proposed conclusions through multiple means.

Robust Big Data Science should be *rigorous*

Rigor is reflected by the extent to which core constructs and variables are operationalized with precision, the methodologies used to gather observations are free from error/bias, data are acquired from samples that are representative and appropriate for desired inferences, and the analytical techniques used to model relationships within data meet required assumptions. Concerns with the rigor of Big Data science are among the most commonly discussed issues in the academic literature, with numerous authors citing the need for Big Data practitioners to more carefully evaluate the quality, appropriateness, and psychometric properties of data used to generate inferences (e.g., Boyd & Crawford, 2013; Braun & Kuljanin, 2015; Guzzo et al., 2015; Hilbert, 2016; Whelan & DuVernet, 2015). Guidance for promoting more rigorous Big Data research are beginning to emerge (e.g., Cai & Zhu, 2015; Landers, Brusso, Cavanaugh, & Collmus, 2016), and we suspect the rigor of Big Data approaches will continue to mature as standards and best practices emerge. Nevertheless, "Big Data hubris" (Lazer et al., 2014) and the failure to scrutinize the rigor of computational social science applications represents a clear threat to promoting a robust Big Data science as they compound the risk of generating inferences that are unreliable, unreproducible, and untrustworthy.

Robust Big Data Science should be *replicated* and *accumulative/cumulative*

Although there are subtle and important distinctions across these two characteristics of robust science, we discuss them collectively in the present treatment as they both speak to the trap of assuming that bigger/more data necessarily implies higher quality inferences (e.g., Boyd & Crawford, 2012; Guzzo et al., 2015; Landers et al., 2016; Lazer et al., 2014). Though many have opined that the replication of findings is the cornerstone of all science (e.g., Simons, 2014), what it means to "replicate" research is a more complicated question than many assume (Anderson & Maxwell, 2016; see also exchange between Gilbert, King, Pettigrew, & Wilson, 2016, and Anderson et al., 2016). The use of relatively small and underpowered sample sizes in psychological research is among the most common reasons why meta-scientists have so strongly advocated for replication studies in the past. While Big Data applications are much less likely to suffer from similar issues of statistical power, random error is not the only potential source of variance that replication efforts may address. Psychologists have long recognized that human affect, behavior, and cognition is responsive to situation, context, time, and myriad other factors that may vary across a set of observations. From this perspective then, even a single Big Data study may still represent an *n* of 1 (albeit a large *n* of 1). Furthermore, and consistent with the significance of rigor to robust science, replicating and/or collecting vast amounts of data using "poorly designed" research (e.g., questionable operationalization of core constructs, use of psychometrically deficient measures, failing to consider the representativeness of a sample for inferences) adds rather than reduces uncertainty around inferences. Consequently, if a critical goal of science is to advance understanding of the natural world, replication should be viewed as efforts to accumulate as many *high quality* observations as possible for a relationship so as to establish the degree of confidence we should place in our cumulative knowledge. Big Data methods can clearly play an important and unique role in helping psychological science

accumulate such knowledge, but it does not preclude the importance of ensuring the reliability and reproducibility of Big Data results through both direct and conceptual replications.

Robust Big Data Science should be *transparent and open*

Transparency and openness in science is most directly embodied by efforts to share and disclose all data, materials, analyses, and hypotheses that comprise a research study with the scientific community (Nosek et al., 2015). There are many platforms available that have made sharing and accessing these items easy for both primary investigators and secondary consumers (e.g., Open Science Framework, https://osf.io; GitHub, https://github.com; Dataverse, https://dataverse.org), and the available features, interconnectivity among, and support for such outlets has continued to increase as more users adopt these technologies. However, fostering a transparent and open science goes beyond sharing data and materials; it also involves concerted efforts to detail the precise processes (i.e., methods/analyses) involved in the procurement and analysis of data rather than only the final outcomes of the research (Grand, Rogelberg, Banks, Landis, & Tonidandel, 2018). This is a particularly important target for promoting a robust Big Data science given that many decisions in the collection, aggregation, processing, wrangling, recording, storing, and analyzing of data can be ambiguous or opaque (Boyd & Crawford, 2012). To this end, we think it likely Big Data science would also benefit from participating in pre-registration, registered reporting, and other similar mechanisms that place greater emphasis on how research questions will be addressed and inferences drawn (Cummings, 2013; Open Science Collaborative, 2015). Even when such options may be unavailable to the researcher, Big Data practitioners should make every effort to accurately and completely document all procedures that impact what goes into and out of analyses and make those data and computations available. Although relevant to all empirical research, we believe this principle carries even greater weight

for Big Data analytics that rely heavily on data rather than theory for insight generation (cf., Landers et al., 2016).

Robust Big Data Science should be *theory-oriented*

A defining feature of science versus other epistemological perspectives is the pursuit of evidence that helps bound, revise, falsify, and advance explanatory claims about the natural world. This definition does not mean that descriptive or correlational research (as might be pursued using unsupervised learning methods) are or should be valued less than research geared towards hypothesis testing or confirmation (as might be pursued using supervised learning methods). Instead, a theory-oriented science is one that builds towards a precise understanding of the magnitude, form, processes, and conditions that account for observed relationships (Edwards & Berry, 2010; Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013). This goal is clearly served by both inductive and deductive perspectives. That said, we believe there is great and untapped potential for Big Data analytics to play a more significant role in efforts to develop and test theory in the psychological sciences than we have seen thus far. There are many examples in which Big Data analytics have demonstrated their unique power to extract intriguing signals from noisy data, but efforts to guide and/or situate this knowledge in the broader context of previous theory often occurs in a more retrospective/abductive manner or not at all. Leveraging the strengths of computational science methods to both generate and evaluate theory would greatly improve the capacity for a robust Big Data science. One area where we envision particularly exciting potential is through the use of Big Data techniques for advancing theory on the dynamic processes that unfold over time within persons and other levels of analysis (e.g., dyads, networks, teams, organizations; Braun & Kuljanin, 2015; Kozlowski et al., 2013; Lazer et

al., 2009; see Kennedy & McComb, 2014, and Kozlowski, Chao, Chang, & Fernandez, 2016 for examples at the team-level).

In sum, creating and sustaining a robust scientific field of inquiry is facilitated when its contributors and stakeholders share similar aspirational values (Grand, Rogelberg, Allen, et al., 2018). Given that Big Data analytics is not a field "owned" or developed by a particular discipline, the users and producers of Big Data research hold a significant responsibility for ensuring that the knowledge generated through these methods is credible, reliable, and relevant. This is likely to be particularly important in the psychological sciences, where we suspect Big Data and computational social science techniques will be intriguing to many but actively pursued and well understood by only a small subset of researchers (Aiken & Hanges, 2015). As a result, the opportunities for self-correction, oversight, and peer evaluation—the traditional safeguards for ensuring scientific integrity—are likely more limited. Consequently, we now direct attention to issues that non-Big Data researchers frequently cite as threats to the reliability and credibility of science that we believe Big Data researchers in psychology should also attend to increase the likelihood their work actively contributes to a more robust psychological science.

Avoiding Pitfalls and Encouraging a Robust Big Data Science in Psychology

There are many stakeholders that contribute to the reliability and credibility of any scientific field (Grand, Rogelberg, Allen, et al., 2018; National Academies of Sciences, Engineering, & Medicine, 2017). However, researchers arguably hold the most central role as the first-line producers, disseminators, reviewers, and consumers of a field's knowledge. We consider three concerns commonly discussed in the broader psychological research literature — hypothesizing after results are known (HARKing), questionable research practices (QRPs), and

replicability/reproducibility—by describing how these concerns might manifest and offering suggestions for minimizing their proliferation in Big Data research.

Hypothesizing after Results are Known (HARKing)

HARKing was originally characterized as the addition or removal of predictions from a research paper once the researcher is aware of the pattern of findings in collected data (Kerr, 1998). This conceptualization has expanded in recent years to more broadly encompass attempts to change and/or redevelop one's hypotheses or proposed theoretical rationale for hypotheses after seeing the results of statistical analyses. For example, a commonly described form of HARKing involves "cherry-picking" statistically significant results and then weaving together (post hoc) a convincing narrative in the introduction to a paper that implies such findings were predicted, consistent with theoretical rationale, and can be packaged into a coherent whole (Banks et al., 2016; Hollenbeck & Wright, 2017).

HARKing holds numerous negative consequences for a scientific discipline. Most notably, the practice can inflate the false positive rate of published findings by increasing the likelihood that the inferential conclusions and claims advanced in a paper are the result of chance or spurious relationships in a study's sample. To be clear, the issue with HARKing is not a statistical one—the presence of a "significant" relationship in a sample does not change based on whether it was predicted a priori. Rather, the concern stems from the philosophy and principles of logic from which the epistemological framework of scientific deduction are rooted. Hypothesis testing implies that a researcher believes a relationship should exist in the natural world on the basis of a theoretical rationale. A methodology is then devised and implemented to gather observations of this relationship that (often) attempts to control or rule out alternative explanations. Finally, the observations are fit to an inferential (i.e., statistical) framework to

evaluate the likelihood of the theoretical claim relative to other claims (e.g., null hypothesis significance testing, interpreting Bayes factors or Bayesian credibility intervals). This process maintains the logical consistency and underpinnings of deductive reasoning (i.e., theory → hypothesis → inference). In contrast, HARKing covertly reverses this process (i.e., inference → hypothesis → theory) and thus undermines the argumentative strength upon which the support for an inferential conclusion and any associated theoretical considerations are derived.

Beyond its direct epistemological concerns for science, HARKing also has the potential for a number of indirect harms (Hollenbeck & Wright, 2017; Kerr, 1998). For example, HARKing can result in theories becoming entrenched in the science that do not actually offer viable causal explanations for relationships. As a result, a field could be misled and its explanatory foundations weakened as others use those spurious claims to generate and integrate new theory. Additionally, valuable researcher time and resources may be expended on efforts to replicate and evaluate the veracity of HARKed findings that emerged through chance variation in a sample. While such replication efforts are warranted and a critical means of correcting such erroneous conclusions, they are regrettable in the case of HARKed results given that the originating research knowingly advanced misguided claims. An even more extreme scenario can be envisioned if one considers that "supportive" results tend to be more frequently published than null results (Fanelli, 2010a, 2011; Ioannidis, 2005a). Thus, papers that end up reproducing HARKed findings—either purposefully or by chance—may be more likely to be published than those that do not, thus further embedding the erroneous inference in the literature.

Relevance and Recommendations for Big Data Research. Given its natural inclination towards quantitative empiricism, the underlying philosophy of Big Data analytics often encourages "data mining" or probing data for unplanned or unanticipated relationships to

generate insights post hoc. For example, a researcher might apply one or more unsupervised learning techniques to identify novel and/or previously unknown groupings in a dataset, leverage various supervised learning techniques to identify potential predictors or covariates of cluster membership, and then produce an interpretation/explanatory rationale for notable relationships (i.e., "generate insight") while ignoring those that appear less promising. In many respects, this process closely resembles the much maligned practice of HARKing described above—particularly if the Big Data researcher subsequently develops a conceptual narrative that neatly fits the particular clusters, predictors, and relationships observed in the dataset and poses it in the introduction of a research paper as the theoretical foundations for the study.

The most critical recommendation for avoiding the pitfalls of HARKing in Big Data research is for the researcher to clearly differentiate which relationships were anticipated a priori from substantive theory and for which the study/methodology was explicitly designed to evaluate from relationships that were unanticipated, observed post hoc, and for which there was there was no explicit intention to infer the veracity of particular theoretical claims (see Hollenbeck & Wright, 2017, for similar conclusions in general psychological research). In cases where one intends to use Big Data methods or techniques in a deductive fashion to evaluate theory-driven hypotheses, the researcher should explicate the conceptual model/rationale for all hypotheses in the introduction section of a paper and evaluate the degree of support for those theoretical claims in the subsequent results and discussion sections (similar to the format commonly used in the majority of published psychological research). In cases where one intends to use Big Data methodologies in an inductive fashion to explore, identify, and discover potential relationships in a dataset, the researcher should communicate the decisions, choices, rationale, and accompanying justification for the way in which data were gathered, processed, and analyzed in

the introduction and methods sections of the paper. The results and discussion sections should subsequently focus on interpreting the generative mechanisms, possible reasons for the observed findings, and their implications for developing new theoretical claims. Lastly, in the (potentially more common) case in which one uses Big Data techniques to both deductively test hypotheses and inductively probe additional and/or alternative relationships, the researcher should clearly delineate these foci in the introduction, methods, and results portions of a paper. The inclusion of one or more sections dedicated to "Exploratory Analysis and Interpretation" is recommended so that consumers and reviewers can easily identify these distinctions and apply the appropriate intellectual skepticism when interpreting and building upon findings from the published work (Grand, Rogelberg, Banks, et al., 2018).

Questionable Research Practices (QRPs)

Although the scientific method and empirical research process is often described as a highly standardized and systematic approach to studying the natural world, anyone who has actually performed research recognizes that it involves a series of judgment calls for which there are often no clearly defined rules of guidelines. For example, deciding how to measure/operationalize constructs; where, when, and how to sample participants; when to terminate data collection; how to treat outliers; or which (if any) control variables to include in an analysis are all decisions under a researcher's control, but for which there is typically no definitive answer. Simmons, Nelson, and Simonsohn (2011) characterize these choice points as "researcher degrees of freedom" and discuss the profound impact they can have on the outcomes and inferences drawn from any given research project. Though researcher degrees of freedom open the door for a wide degree of variability in scientific practices, they are largely unavoidable and not inherently threats to the robustness of science (e.g., McGrath, 1982). However, they

become clear causes for concern when researchers make such judgment calls in ways that present more favorable evidence for proposed claims or hypotheses. Decisions and associated actions of this nature are commonly referred to as questionable research practices (QRPs; e.g., Banks et al., 2016).

It should be noted that all but the most egregious (e.g., falsifying or fabricating data) of questionable research practices is likely not motivated by malicious intent of a researcher to disseminate misinformation or misrepresent their data/claims. Simmons et al. (2011) intimate that participation in QRPs is more likely a result of ambiguity in how to resolve researcher degrees of freedom coupled with the researcher's hope/optimism of finding results that support their hypotheses. Others also cite the influence of norms within academia (e.g., "publish or perish") and the broader research enterprise (e.g., journal criteria that emphasize "novel" or "counter-intuitive" findings) as relevant contributing factors (e.g., Anderson, 2007; Fanelli, 2010b; Rawat & Meena, 2014). Irrespective of the cause, the negative implications of QRPs are clear and similar to those described for HARKing (e.g., inflated false positive rate of published findings, erroneous inferential conclusions, ambiguous theoretical evidence). Furthermore, and potentially more insidious, is the precedent that QRPs set for future methodological practices. Whereas HARKing encourages "creative story-telling" and the development of post hoc rationalizations for observed findings, engaging in QRPs additionally encourages "playing" with data and statistics until desired relationships are found.

Relevance and Recommendations for Big Data Research. Research utilizing Big Data methodologies is likely to be just as susceptible to QRPs as any other form of research. However, the specific ways in which they manifest may differ. Additionally, the techniques, standards, and affordances related to collecting large datasets (e.g., web scraping, trace data, wearable sensors)

and the analytic practices used to model such data are still emerging and further contribute to uncertainties regarding how to resolve researcher degrees of freedom in Big Data applications.

With respect to data gathering/collection, subsetting (removing/focusing on specific observations from a dataset on the basis of particular characteristics) and fusing (complementing data with other properties of the data source or observations from a secondary data source) data are common practices in Big Data research (Cheung & Jak, 2016; Hilbert, 2016; Braun & Kuljanin, 2015). For example, suppose a researcher is interested in studying political echo chambers and predicts that the ideology of journalists' social media networks and the news content they produce are correlated (e.g., those whom follow more liberal (conservative) Twitter accounts write more liberal (conservative) articles).[1] Upon gathering their initial dataset of over 500,000 articles from 1000 journalists, the researchers find that the results seem "close" to supporting their prediction, but there are a number of unusual observations (e.g., highly prolific authors, small number of Twitter follows). The researchers elect to remove those data points from the analysis and observe that the subsetted dataset—which still includes roughly 300,000 articles from 750 journalists—brings the findings directly in line with the study's predictions. The reduced dataset is subsequently reported in the final research product with little discussion of outlier removal or its effects on the study interpretations. Alternatively, suppose the authors find that operationalizing the ideology of a person's social media networks using only Twitter data does not reveal the predicted trend. However, the hypothesis is supported when that data is combined with a dataset tracking journalists' Facebook "likes" on political events. The composite metric is subsequently used for the researchers' analyses and is the only one reported

[1] This example is based on a study by Wihbey, Coleman, Joseph, and Lazer (2017). We make no claims about the methodological practices or presence of potential QRPs in this work and only use the research question explored by the authors for pedagogical purposes.

in the final research product. While either decision seems innocuous and arguments defending those choices could be made, such decisions venture dangerously close to the territory of QRPs as they make it difficult to faithfully evaluate the generalizability (in the case of data subsetting) and validity (in the case of data fusion) of the final inferences.

Related QRPs can emerge in a variety of Big Data analytic techniques as well. For example, most machine learning techniques depend heavily on the quality of the training samples used to inform their prediction and classification routines (Oswald & Putka, 2015). Consequently, different sizes or compositions within one's hold-out samples could result in different patterns of results that could be knowingly or unknowingly leveraged to support particular claims. Additionally, many analyses contain parameters that can be "tweaked" to produce different results and thus represent additional researcher degrees of freedom. For example, determining the number of nodes, layers, and number of connections among nodes in artificial neural networks is described by some as more "art than science" and can result in different conclusions about predictor-criterion relationships (e.g., Jain & Mao, 1996). Similarly, decision tree/random forest models can be adjusted to account for more global versus local optimization of predictors (Oswald & Putka, 2015). In both cases, the results produced under different configurations are no less "correct" than other parameterizations and can be explored to identify conditions under which particular inferences/conclusions are justifiable. Although we suspect that criteria or rules of thumb for such choices are likely to emerge as Big Data analyses continue to mature, a large majority of psychologists will not possess the requisite expertise to evaluate the significance of such choices (Aiken & Hanges, 2015) and thus detecting these potential QRPs will remain difficult.

In many respects, recommendations for limiting the impact and prevalence of QRPs in Big Data research are similar to those proposed for improving the robustness of more general psychological research. These suggestions largely revolve around improving the transparency with which researcher degrees of freedom are resolved and journal reporting standards that necessitate these discussions in published materials (e.g., Simmons et al., 2011; Nosek et al., 2015). In addition to these practices, we also encourage Big Data researchers to explore the use of pre-registration and alternative publication mechanisms (e.g., registered reports, results-blind reviews; Grand, Rogelberg, Banks, et al., 2018; Cummings, 2013; Open Science Collaborative, 2015). Though it may be difficult to adapt Big Data research that is more inductive and exploratory in nature to use these avenues, encouraging researchers to carefully and explicitly consider their measurement, data gathering, and analysis plans prior to collecting or observing any data should be beneficial. For example, explicitly declaring what, how much, and from where data will be collected can help prevent the introduction of QRPs once a study is underway. Any deviations from this plan can still be made so as not to stifle creativity and innovation; the key is simply to be transparent, explain where any deviations occur, and why they are justifiable.

On a related note, we suggest Big Data researchers that conduct research utilizing archival or scraped data adhere to Landers et al.'s (2016) recommendation to construct and refer to a *data source theory* prior to, while carrying out, and when reporting their research. A data source theory describes the assumptions a researcher must make about a prospective data source to be able to extract meaningful inferences from it. For example, describing why individuals create the available data and what it represents (e.g., Do Facebook "likes" indicate agreement with or acknowledgement of the original content?), which individuals have access to and are likely to participate in the data source (e.g., Is website for seeking social support or sharing

opinions?), and where/how data are structured (e.g., Is information located on private profile pages versus open forums?) facilitate understanding of the data's context as well as information upon which to base interpretations of construct validity (Boyd & Crawford, 2013; Braun & Kuljanin, 2015; Guzzo et al., 2015; Hilbert, 2016; Whelan & DuVernet, 2015). Data source theories can also be used to help elaborate why particular data collection or analytic strategies were *not* chosen or were deemed less desirable. In sum, the goal of practices to combat QRPs in Big Data research is to provide information that allows readers/reviewers to more accurately determine the extent to which a study's inferences hinge on the researcher's methodological and statistical decisions or whether the conclusions are robust to alternative operationalizations.

Replicability and Reproducibility

As noted previously, many have expressed concerns over the replicability and reproducibility of research across the sciences in general and the psychological sciences in particular. Although the terms are often used interchangeably, replication and reproduction should be differentiated in discussions of robust science as they carry different implications for establishing confidence in scientific results (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015). Furthermore, efforts to examine the replicability versus reproducibility of scientific claims often rely upon different methodologies and sources of evidence to judge. It is thus useful to distinguish between these concepts when considering their significance in Big Data applications.

Replicability is generally defined as the capacity to "duplicate the results of a prior study if the same procedures are followed but new data are collected" (Bollen et al., 2015, p. 4). In this sense, efforts to replicate research findings most often involve determining whether similar findings, conclusions, and interpretations presented in an *existing* study can be observed in a *new* study. However, Anderson and Maxwell (2016) note that replication studies can serve many

purposes and highlight the accompanying methodological/analytical strategies and criterion for evaluating replication "success." For example, a replication study could be carried out to infer whether an observed effect exists, in which case the replication researcher should attempt to conduct the exact analyses of the original study and evaluate whether the new effect is in the same direction and significant. Alternatively, a replication study could be performed to assess whether a replication is inconsistent with the original observation, in which case the researcher should rely on evaluating confidence intervals of effect size differences. Other researchers have also debated the merits and differences between direct replications (e.g., efforts to reach the same conclusion using identical procedures) versus conceptual replications (e.g., efforts to reach the same conclusion using different procedures; Simons, 2014; Stroebe & Strack, 2014). Irrespective of its form, though, the focus of most replication efforts concerns establishing the generalizability of inferential claims (i.e., how much confidence should be placed in the veracity and robustness of an empirical conclusion or theoretical claim).

In contrast, reproducibility is usually defined as the capacity to "duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" (Bollen et al., 2015, p. 3). Reproducing research findings thus typically involves determining whether the findings, conclusions, and interpretations presented in an *existing* study can be recreated using the materials, data, and analyses from the *same* study. Goodman, Fanelli, and Ioannidis (2016) expanded upon this definition to differentiate three considerations of reproducibility. The first, *methods reproducibility*, considers whether the procedures, steps, and choices used in the original study can be exactly repeated/reconstructed and is typically evaluated by the extent to which the data collection/measurement process, data processing, and analytical reporting are sufficiently detailed in a published product. Goodman et al. (2014)

equate their second form of reproducibility, *results reproducibility*, with direct replication. However, this criterion can also represent the degree to which an independent researcher can use the same raw data to run the same statistical analysis and produce the same statistical result presented in a published research study (Bollen et al., 2015). Lastly, *inferential reproducibility* describes whether the same interpretations drawn by an original study can be reached under different assumptions about the data, statistical models used, or evaluative criteria. In sum, the primary focus of reproducibility evaluations tends to concern the internal validity of the procedures used to generate inferential claims (i.e., can the methodological steps be determined, are they robust to different researcher decisions, and do they lead to the stated conclusions).

Relevance and Recommendations for Big Data Research. Although both replicability and reproducibility provide value to establishing the robustness of scientific research, they also offer unique costs and benefits that hold important implications for Big Data researchers. Many scientists acknowledge that the direct replication of results is the definitive standard and reproduction of results a minimal standard for establishing the veracity of scientific claims (e.g., Bollen et al., 2015; Peng, 2011; Sandve, Nekrutenko, Taylor, & Hovig, 2013; Simon, 2014). However, replication efforts are usually much more time- and resource-intensive as they require a researcher to acquire new data sources that are equated on as many methodological factors with the original study as possible (e.g., operationalization and measurement of critical constructs, sample and contextual characteristics, etc.). This demand may render direct replication of Big Data research infeasible at best, or near impossible at worst. Further, some data scientists have argued that direct replication efforts may be antithetical to the inherent strengths and uses that Big Data techniques offer for generating unique insights (e.g., Drummond, 2009).

Though efforts to directly replicate Big Data research face a litany of unique and difficult complications, we believe there are important actions that Big Data researchers can take towards this end. Specifically, we recommend that Big Data researchers participate in data sharing and open science practices to facilitate cumulative verification of results. As Anderson and Maxwell (2016) note, the goals of replication can be much broader than attempts to conclude whether the findings from a particular study hold. Instead, replication efforts can contribute to the broader goal of improving the veracity and degree of confidence a field should place in generated knowledge (Grand, Rogelberg, Allen, et al., 2018). Poldrack and Gorgolewski (2014) summarize how this philosophy is being adopted by a growing number of neuroscientists and the subsequent development of open access repositories for sharing neural imaging data. As more researchers contribute to these repositories—the majority of whom are not attempting to directly replicate a particular study but have gathered observations of relevant phenomenon—the capacity to evaluate/verify brain activation differences among healthy controls and individuals with clinical diagnoses (in a true or quasi-Bayesian manner) has improved. We suspect that the construction, maintenance, and regulation of such repositories will be idiosyncratic for some time as more individuals begin to dabble in Big Data techniques. However, good models (such as those cited by Poldrack & Gorgolewski, 2014) and first principles for how to manage and grow these repositories are becoming available and only expected to improve in the coming years.

Though not without its own challenges, improving the methods, results, and inferential reproducibility of Big Data research appears to be a more attainable and oft discussed goal by Big Data researchers. The most common recommendation discussed for achieving this standard involves improving the documentation, disclosure, and dissemination of the procedures used in the collection, processing, and analysis of Big Data research. For example, Sandve and

colleagues (2013) offer ten rules for improving the reproducibility of computational science, with suggestions ranging from maintaining accurate version control of scripts and analytical software to keeping records of how results were produced from start to finish. Peng (2011) also describes a useful heuristic dubbed the "reproducibility spectrum" that provides authors, reviewers, and editors a classification system for characterizing and suggesting ways to enhance the reproducibility of published research products. At the lowest end of this spectrum is the basic journal article in which methods and analyses are described in the manuscript text, appendices, and/or other supplemental documents. From there, reproducibility can be increasingly improved by providing (1) all computer code used to process, analyze, and/or produce data; (2) all code plus all the raw data used in the reported analyses; and (3) fully executable code and data that links directly to the conclusions and inferences produced in the manuscript. We also recommend that demonstrating and/or including the means to more readily evaluate the robustness of inferential conclusions to alternative analytic parameterizations as part of submitted code would make a useful and positive contribution to the reproducibility of Big Data research.

Conclusion

The use and potential of Big Data and computational methodologies for furthering psychological research on human affect, behavior, cognition, and relationships is both thought-provoking and energizing. The impetus of this chapter was motivated by considering what Big Data and computational social scientists could do to improve the likelihood that its research meets emerging criteria for robust and reliable psychological science. In many ways, it is serendipitous that interest in Big Data techniques has coincided with the most recent surge of calls to action for safeguarding the credibility and trustworthiness of scientific research. The psychological sciences have taken their share of bumps and bruises in this domain. However,

new norms, standards of practice, and tools for countering potential threats to the robustness of psychological research are emerging more rapidly than ever before and are incrementally changing how research is performed, reviewed, and disseminated. We believe these developments are equally applicable and critical to the burgeoning areas of computational social science, and hope that conversations regarding how to promote robust and reliable Big Data research in psychology continue to unfold.

References

Aiken, J.R., & Hanges, P.J. (2015). Teach an I-O to fish: Integrating data science into I-O graduate education. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8,* 539-544.

Anderson, C.J., Bahnik, S., Barnett-Cowan, M., Bosco, F.A., Chandler, J., Chartier, C.R., … Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037-c.

Anderson, S.F., & Maxwell, S.E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1-12.

Banks, G.C., O'Boyle, E.H., Jr., Pollack, J.M., White, C.D., Batchelor, J.H., Whelpley, C.E., Abston, K.A., Bennett, A.A., & Adkins, C.L. (2016). Questions about questionable research practices in field of management: A guest commentary. *Journal of Management, 42,* 5-20.

Bollen, K., Cacioppo, J.T., Kaplan, R.M., Krosnick, J.A., & Olds, J.L. (2015). *Social, behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.* Arlington, VA: National Science Foundation.

Boyd, D., & Crawfrod, K. (2012). Critical questions for Big Data. *Information, Communication, & Society, 15,* 662-679.

Braun, M.T., & Kuljanin, G. (2015). Big Data and the challenge of construct validity. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8,* 521-526.

Cai, L., & Zhu, Y. (2015) The challenges of data quality and data quality assessment in the Big Data era. *Data Science Journal, 14,* 1-10.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2,* 637-644.

Cheung, M., & Jak, S. (2016). Analyzing Big Data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7,* 1-13.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Edwards, J.R., & Berry, J.W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668-689.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. In L. Bottou & M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning*, Montreal, CA. Retrieved from: http://cogprints.org/7691/7/ICMLws09.pdf.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE, 4,* e5738.

Fanelli, D. (2010a). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*, e10068.

Fanelli, D. (2010b). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE, 5*, e10271.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891-904.

Gilbert, D.T., King, G., Pettigrew, S., & Wilson, T.D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351*, 1037-b.

Grand, J.A., Rogelberg, S.G., Allen, T.D., Landis, R.S., Reynolds, D., Scott, J.C., Tonidandel, S., & Truxillo, D.M. (2018). A systems-based approach to fostering robust science in Industrial-Organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*, 4-42.

Grand, J.A., Rogelberg, S.G., Banks, G.C., Landis, R.S., & Tonidandel, S. (2018). From outcome to process focus: Fostering a more robust psychological science through registered reports and results-blind reviewing. *Perspectives on Psychological Science, 13*, 448-456.

Grote, G. (2016). There is hope for better science. *European Journal of Work and Organizational Psychology, 26*, 1-3.

Guzzo, R.A., Fink, A.A., King, E., Tonidandel, S., & Landis, R.S. (2015). Big Data recommendations for Industrial-Organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*, 491-508.

Hilbert, M. (2016). Big Data for development: A review of promises and challenges. *Development Policy Review, 34*, 135-174.

Ioannidis, J.P.A. (2005a). Why most published research findings are false. *PLoS Medicine, 2*, 0696-0701.

Ioannidis, J.P.A. (2005b). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*, 218-228.

Jain, A.K., Mao, J., & Mohiuddin, K.M. (1996). Artificial neural networks: A tutorial. *Computer, 29*, 31-44.

Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: Insights from simulation and optimization. *Journal of Applied Psychology, 99*, 784-815

Kim, G.H., Trimi, S., & Chung, J.H. (2014). Big Data applications in the government sector. *Communications of the ACM, 57*, 78-85.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams Jr, R.B., Bahník, Š., Bernstein, M.J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology, 45*, 142-152.

Kozlowski, S. W. J., Chao, G. T., Chang, C.-H., & Fernandez, R. (2015). Team dynamics: Using "big data" to advance the science of team effectiveness. In S. Tonidandel, E. King, & J. Cortina (Eds.), Big data at work: The data science revolution and organizational psychology (pp. 272-309). New York, NY: Routledge Academic.

Kozlowski, S.W.J., Chao, G.T., Grand, J.A., Braun, M.T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods, 16*, 581-615.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods, 21*, 475-492.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in Big Data analytics. *Science, 343*, 1203-1205.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, *323*, 721-723.

Leetaru, K. (2017). A case study in Big Data and the replication crisis. *Forbes*. Retrieved from: https://www.forbes.com/sites/kalevleetaru/2017/09/01/a-case-study-in-big-data-and-the-replication-crisis/#281f1c045105.

Marcus, E. (2014). Credibility and reproducibility. *Cell, 159*, 965-966.

McGrath, J.E. (1982). Dilemmatics: The study of research choices and dilemmas. In J.E. McGrath, J. Martin, & R.A. Kulka (eds.), *Judgment calls in research* (pp. 69-102). Beverly Hills, CA: Sage.

National Academies of Sciences, Engineering, & Medicine. (2017). *Fostering Integrity in Research*. Washington, DC: The National Academies Press

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., … Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348,* 1422-1425.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

Peng, R.D. (2011). Reproducible research in computational science. *Science, 334,* 1226-1227.

Poldrack, R.A., & Gorgolewski, K.J. (2014). Making big data open: Data sharing in neuroimaging. *Nature Neuroscience, 17,* 1510-1517.

Rawat, S., & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences, 19,* 87-89.

Rubin, M. B. (2011, July). Fraud in organic chemistry. *Chemistry in New Zealand, 78,* 128-132.

Ryan, J., & Herleman, H. (2015). A Big Data platform for workforce analytics. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 19-42). NewYork, NY: Routledge.

Sandve, G.K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PloS Computational Biology, 9,* e1003285.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76 – 80

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.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59-71.

Whelan, T.J., & DuVernet, A.M. (2015). The big duplicity of Big Data. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8,* 509-515.

Whiby, J., Coleman, T.D., Joseph, K., & Lazer, D. (2017). Exploring the ideological nature of journalists' social networks on Twitter and associations with news story content. *arXiv preprint (https://arxiv.org/abs/1708.06727v2).*

Table 1.
Defining principles of robust science and their implications for Big Data research

Robust science is…	Description	Implications for Big Data research
Relevant	Generation and application of research is intended to improve understanding of the natural world, address contemporary needs and issues, and/or contribute to beneficial societal outcomes	• Goals, purpose, and focus of research are made explicit • Exploratory/insight-driven research is presented as such
Rigorous	Theoretical and empirical activities emphasize careful operationalization of core concepts and use of diverse methodological/analytical approaches to explore research questions	• Consideration is directed towards validity, reliability, and psychometric properties of data • Appropriateness (rather than size) of data source for examining relationships is justified • Big Data used in addition to, rather than replacement for, existing methodologies
Replicated	Collection of multiple and repeated observations of primary relationships are pursued and recognized as critical to establishing confidence in scientific claims and evidence-based practice	• Sensitivity of inferences to alternative models and specifications is examined and reported • Examining relationships described in previous data is valued and pursued with new data sources
Accumulative and cumulative	Cumulative knowledge and efforts to establish confidence in the strength of scientific understanding are pursued in a manner that balances generation and incremental vetting of new ideas	• Relationships identified in previous data are integrated into/accounted for in new data • Big Data approaches used in both confirmatory and exploratory manners
Transparent and open	Activities related to conducting, reporting, and disseminating research are undertaken in ways that facilitate understanding of the processes involved and products created during research	• Data sources, methods, and analyses are shared • All data processing, wrangling, and recording decisions are shared • Participate in registered reporting, pre-registration, and other mechanisms that emphasize research process
Theory-oriented	Outputs of all scientific research contribute to the development of increasingly accurate, useful, evidence-based, and precise explanations for natural phenomena observed in the world	• Big Data used to bound, revise, and falsify in addition to advancing new claims • Big Data used to improve precision of process-level accounts for phenomena

Note. Table adapted from Grand, J.A., Rogelberg, S.G., Allen, T.D., Landis, R.S., Reynolds, D., Scott, J.C., Tonidandel, S., & Truxillo, D.M. (2018). A systems-based approach to fostering robust science in Industrial-Organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*, 4-42, and distributed under the Creative Commons Attributions 4.0 License