Protect Yourself from p-Hacking: 7 Things to Do to Avoid Committing Scientific Malpractice

Frederick Solt
frederick-solt@uiowa.edu

Yue Hu
yue-hu-1@uiowa.edu

Kevan Hudson
kevan-hudson@uiowa.edu

Jungmin Song
jungmin-song@uiowa.edu

Dong ‘Erico’ Yu
dong-yu@uiowa.edu

December 25, 2015

Abstract

Replication crisis

LaCour scandal

p-Hacking. check out special issue on p-hacking: http://www.tandfonline.com/toc/hcms20/9/4

malpractice not always so blatent or intentional: confirmation bias, garden of forking paths (see http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf)

Introduce Newman, Johnston, and Lown (2015*a*), perhaps noting the press attention it has received (e.g., http://www.psmag.com/health-and-behavior/five-studies-bernie-sanders-says-the-rich-are-deranged)

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=9911378&fulltextType=LT&fileId=S2049847015000448

Leek and Peng, "Reproducible Research Can Still Be Wrong: Adopting a Prevention Approach" http://arxiv.org/abs/1502.03169: "Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. While improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated, it does not change the fact that problematic research is conducted in the first place." (p1)

Sound research is a cornerstone of the advancement of knowledge across all fields of study. The need for carefully conducted, replicable empirical research is not a new concern, though it has gained additional attention of late. The rise of p-hacking, the practice of manipulating data and methodological approaches in the aim of generating statistically significant results, fuels this growing concern. Recent instances highlight the growing concern. Uncovered by the Collaboration (2015), the replication crisis highlights the severity of the problem. Of 100 published studies that the authors attempted to replicate, just 36 produced the same results as those cited in the published articles (Collaboration 2015). Additionally cases such as that of Michael LaCour, the UCLA graduate student who fabricated data, provide further proof of the need for vigilance in combatting academic fraud and malpractice. Given the pressure to produce publishable work, careful consideration of how to detect and minimize academic fraud, along with processes to avoid engaging in questionable research practices are of the utmost importance.

It is important to note that academic misconduct and faulty research need not be the product of an intentional attempt to dupe the system. Confirmation bias, a well-documented tendency to accept those pieces of evidence that confirm our pre-existing beliefs while discounting evidence that counters them undoubtedly can contribute to questionable research practices. Results that support a carefully crafted theory may be accepted as robust, even without thorough examination, while those results that disconfirm the same theory may be dismissed as misspecification. Even the process of developing a model aimed at testing a theory might contribute to faulty research. **?** analogy of a garden of forking paths suggests that even when researchers are not actively engaged in fishing for results, decisions that are made about how to carry out the research necessarily preclude other comparisons from being considered. These decisions, if not carefully considered, can lead to unintended malpractice.

Our aim in this paper is to provide a guide to avoiding academic malpractice through a series of steps, that if followed, are sure to produce sound results that are replicable and transparent, a goal that if accepted widely across academia, could minimize the occurrence of academic fraud and halt the replication crisis in its tracks. In order to demonstrate the importance of these steps, and how issues can arise when they are ignored, we utilize a recently published article, “False Consciousness or Class Awareness? Local Income Inequality, Personal Economic Position, and Belief in American Meritocracy” by Newman, Johnston, and Lown (2015*a*).

This article, published in the American Journal of Political Science examines the effect of local inequality on individual’s belief in the meritocratic nature of America. The findings suggest that rich and poor respond differently to local inequality, with the rich being no less likely to reject meritocracy when inequality is high, while higher inequality increases the likelihood the poor will reject meritocracy. Seen as providing further evidence of the growing gulf between the rich and poor, not just in terms of economic resources, but in their perceptions of the system they cohabitate, the results have gained media attention beyond the field of political science, being featured in an article in Pacific Standard which summarizes a growing body of evidence documenting the widening psychological gap between the rich and poor.

However, the article’s empirical results are misinterpreted, and efforts to replicate its analyses reveal additional problems. As a result, its sanguine conclusions regarding the prospects for redistribution are unsupported. As a result, Newman, Johnston, and Lown (2015*a*) provides a cautionary tale that highlights seven steps that can help researchers to avoid committing academic malpractice, both of the unintentional and intentional sort. Below we outline each of these seven steps, ranging from ensuring that results are replicable, to exercising care in coding and imputing missing values, to properly interpreting interactice effects. We rely on Newman, Johnston, and Lown (2015*a*) to provide illustrations of the dangers that can arise when researchers fail to adhere to these simple processes in the hopes that the mistakes made by its authors might help prevent others from falling into similar situations.

1 Ensure Reproducibility

What is reproducibility? Reproducibility means that Researcher B obtains exactly the same results that were originally reported by Researcher A (e.g. the author of that paper) from A's data when following the same methodology (Brunswik 1955; Asendorpf et al. 2013). Even though it seems like an unimportant process, certainly it is not. Reproducibility is "the gold standard for scientific research" (Janz 2015, 1) because replicating existing studies is "the only way to understand and evaluate an empirical analysis fully" (King 1995, 444). This is why the American Political Science Association (APSA) implemented new guidelines for replication that requires researchers to provide enough information about how they collected the data they used and how they used data to arrive at the published results (Lupia and Elman 2014).

To be specific, ensuring reproducibility can help our discipline in multiple ways. First, without reproducibility, it is impossible to determine the strength or robustness of published results (Lupia and Elman 2014). As the LaCour Scandal in 2015 revealed, the advanced method, the rigorous experimentation process, and the authority of the famous academic journal themselves do not guarantee the reliability of findings. Only after a couple of researchers in the same field attempted to replicate LaCour and Green's findings and failed to do so, our discipline could find out that their findings are completely fraud (Young and Janz 2015). As Dafoe (2014, 62) argues, "Fragile, misleading, and nonreplicable statistical analyses can be largely eliminated" by the replication process.

Second, reproducibility ensures that the future researchers can enjoy “all the benefits of the first researcher’s hard work” (King 1995, 445). Political science has developed as the collective works of numerous researchers. As we can see in the term “community enterprise”, researchers can develop their ideas and methods by extending high-quality existing research (King 1995). However, without enough information about the existing research, the process of collective works cannot continue. Compared to other disciplines, political science requires researchers to spend more time for deciding how to measure and quantify real world observations and events. In this sense, when scholars fail to document enough information about how they collected and used data for results, the published article cannot be fully understandable and effectively interpretable by other scholars (Lupia and Elman 2010).

However, in spite of the shared consensus about the importance of reproducibility and the announced APSA guidelines on data access and research transparency, reproducibility is still a difficult mission to complete. Scholars often do not want to share their data because they have invested a great deal of time and other resources into collecting data and want to get benefits from their data as many as possible (Lupia and Elman 2014). As the bare minimum for replication, now researchers share their replication data in repositories such as the Inter-university Consortium for Political and Social Research (ICPSR), the Dataverse Network at Harvard University, and other journal-specific archives, but we find out that even the replication data in these repositories cannot guarantee the reproducibility of their findings. Newman, Johnston, and Lown’s (2015*b*) replication materials about “False Consciousness or Class Awareness? Local Income Inequality, Personal Economic Position, and Belief in American Meritocracy” do not lead us to obtain the same results as the original paper. They shared just simple codes for table results without the detailed information about how their data is collected and used. Furthermore, this replication paper finds out that the results of Table 1, 2, and 3 of the original paper are not reproducible even with their codes and data shared in the Dataverse Network. In their replication R code, they admit that “the coefficients (of this replication file) will therefore be slightly different, but the signs, significance, and effect sizes remain the same.” If we cannot get the same results with this data, how can we call this data as the “replication data”? As we will show below, the results of Table 1 and 2 are not reproducible exactly, and the result of Table 3 is not reproducible at all because there are more parameters than observations.

Reproducibility as bare minimum for replication; DA-RT APSA guidelines

script all work

packrat, checkpoint, switchr, and pkgsnap packages in R; version command in Stata

2 Work in Public

Researching transparency is always an important criterion for scientific researches, including qualitative and quantitative political scientist studies (Appadurai 2000; Denzin and Lincoln 2009). One way to achieve it is to open every data managing and estimating step in public. It may sound time consuming, unsafe, or unnecessary, especially in the views of researchers who already provide "replication file." Nevertheless, the issue is not only that the replication files are not always replicable in social scientific studies (Chang, Li et al. 2015; Jacoby 2015; Collaboration 2015). But also for the readers who intend to really replicate some part or the entire analysis for further studies, the "replication file" in many cases may not be adequate to fully track the decisions and manipulations the authors made. In such cases, a public-accessed file provides another chance to review what exactly happened during the analyses and with what decisions and movements the conclusion was conducted.

Nowadays, the contemporary computer and internet technologies offers easy and safe ways for researchers to work in public. Taking GitHub as an example,[1] what requires the researcher to do is simply to spend 5 second to build a repository for later pushing updated committed file after changes, as we did for this project. This way also benefit the researchers by recording every step during the analysis, in order to later review, restore, and create replicate files. At the same time, the study per se is also safe as long as the paper and data is not accessed.

Another way to do publicly study is to preregister the research. It is in the same line as pushing staged analyses onto GitHub, but under a more specific supervision of the academia. The preregistration asks researchers to publish their research plan prior to conducting the analyses. The purpose of this is mainly to avoid the result-oriented researches in which the

[1] There are alternative agents, such as SVN and Gitlab, offering similar services, although there might be slight difference in operations and compatible software, and most of them are free to register and use.

researcher manipulate the data or change the theory and hypotheses based on the empirical results they have.[2]

3 Examine All Available Data

examine as much relevant evidence as possible

discuss Figure 1

discuss Figure 2

An important step in the process of conducting sound research is striving to utilize all available data. Researchers should seek to conduct their analysis on the entirety of data that they have at their disposal. Doing so provides a number of benefits: it increases the likelihood that the sample utilized captures the true distribution of the underlying population, and it affords greater leverage in testing the implications of one's hypothesis. While the issue of selection bias cannot be avoided simply by including all available data, limiting analysis to a particular dataset, particularly when alternatives are available, may cast doubt on the inferences drawn from that analysis. By including all relevant data researchers are better able to observe the implications of their theory, thus providing greater support for the hypotheses they advance.

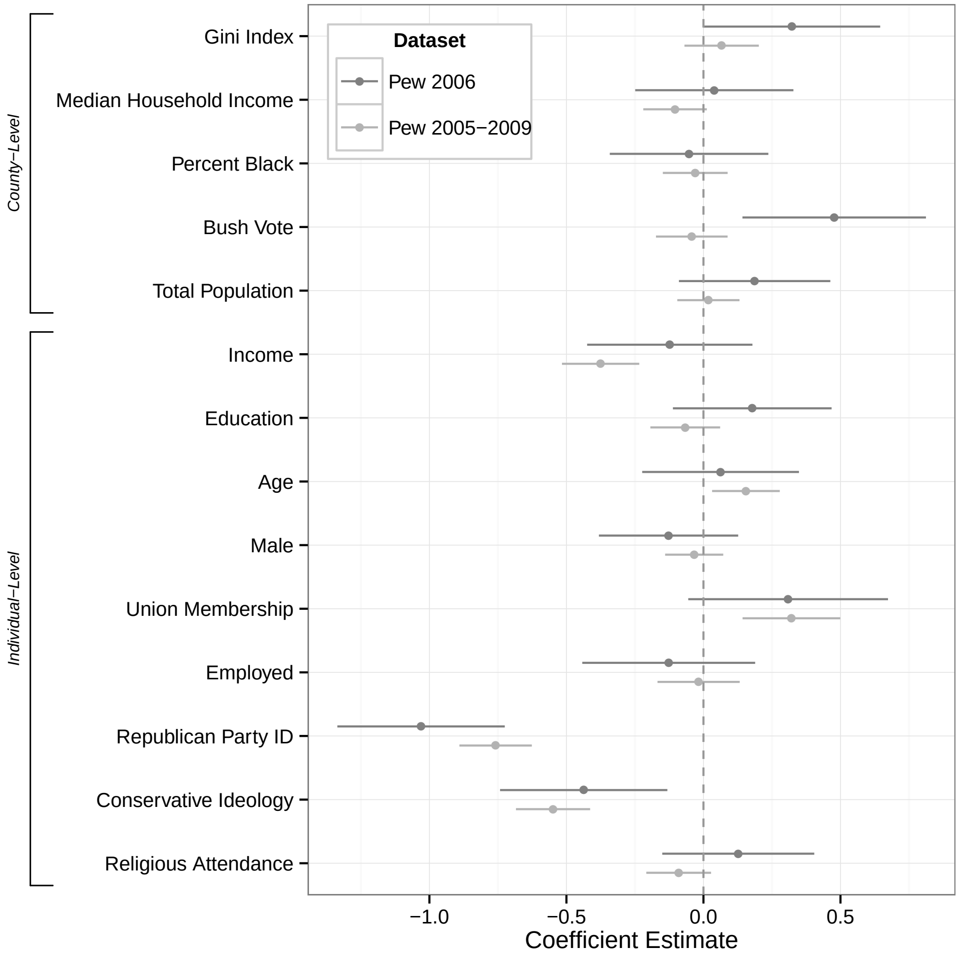
If research is limited to a particular source of data, the findings may be called into question. The limited data may provide evidence of a relationship that is not present in a larger, more representative sample. We can draw an example of the dangers of not including all relevant data from Newman, Johnston, and Lown (2015*a*). In an attempt to test some underlying assumptions of their theory the authors rely on 2006 Pew Research Center dataset due to its "unique set of questions tapping perceptions of economic hierarchy and inequality and respondents perception of their own position within such a hierarchy" (Newman, Johnston, and Lown 2015*a*, 336). As presented by the authors, this dataset provides the only source of information on responses to the question whether or not an American thinks of America as being divided into haves and have-nots, and whether they think of themselves as being haves or have-nots. Employing this dataset the authors find further support for their theory, in situations of higher inequality respondents are more likely to believe that America is divided in such a way, and the poor are more likely

[2] More discussions about preregistration are in the "Symposium on Research Registration", of Winter 2013 issue of *Political Analysis*

to identify themselves as have-nots. In reality, these questions are not unique to the 2006 dataset, but are instead present in each of the surveys the authors used in their earlier analysis. Perhaps it is by coincidence that, as shown in Figure 2, the coefficient of interest only achieves statistical significance when using the 2006 data. As illustrated, no other dataset produces a statistically significant coefficient for Gini according to the authors' model. This provides a clear illustration of the importance of including all relevant data; failure to do so can lead to biased results.

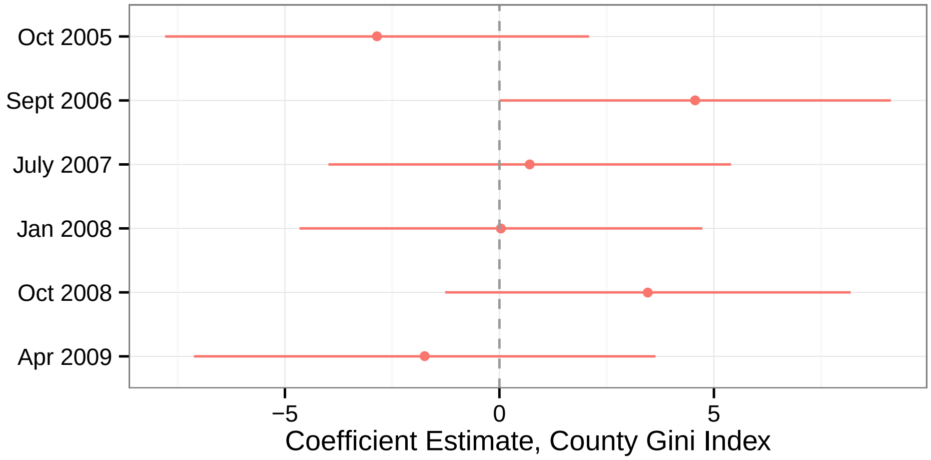
The authors' use of this severely truncated data has implications beyond the coefficient of interest as well. While Figure (INSERT FIGURE NUMBER) clearly demonstrates that a more careful inclusion of all data produce results that run counter to the findings of Newman et al, including all available data drastically changes the entre model, not merely the coefficient for Gini. Figure 1 provides estimates of the coefficients from Newman et al's Table 2 (p. 336) with a sample that includes data from the 2005, 2006, 2007, and 2009 surveys they use earlier in their article. When all relevant data is included, the results are drastically different. Not only does the primary variable of interest (Gini coefficient) lose statistical significance, but others do as well. Having voted for Bush is no longer a statistically significant predictor of believing America is divided into the haves and have-nots, but income becomes strongly negative and significantly associated with the same belief. Additionally, union membership gains statistical significance, indicating that belonging to a union increases the likelihood an individual perceives that have/have-not division. Ultimately, Figure (INSERT NUMBER) provides graphical representation of the dangers of not including all available data. By limiting their analysis to the sole dataset that produced a statistically significant coefficient for inequality, the authors have disguised the true relationship in order to support their theory. A properly crafted analysis reveals findings that are far less surprising; the wealthy are less likely to see America as divided into the haves and have-nots, while union members are more likely to do so. These findings counter the primary argument advanced by Newman et al, and lend strong evidence to the claim that "we should be willing to take whatever information we can acquire so long as it helps us learn about the veracity of our theory," while illustrating the pitfall of picking and choosing data that confirm our theory, while ignoring data that does not (King, Keohane, and Verba 1994, 31).

Figure 1: Local Inequality and the Perception of America as Divided into 'Haves' and 'Have-Nots': Results Using All Available Data



Notes: Results from replications of the model presented in Table 2 of Newman, Johnston, and Lown (2015*a*) on the 2006 Pew survey analyzed in that article and on pooled data from the six Pew surveys that included the same item and were conducted in the time period the article examines. The statistically significant result for county income inequality in the 2006 survey presented in that article is not evident when all of the available data are examined.

Figure 2: Local Inequality and the Perception of America as Divided into 'Haves' and 'Have-Nots': Results Using Each Available Dataset



Notes: Results for county income inequality from replications of the model presented in Table 2 of Newman, Johnston, and Lown (2015*a*) on data from each of six available surveys conducted in the in the time period examined in that article. Of the six surveys, the only one that yields a statistically significant result is the 2006 survey presented in the article.

4 Use Consistent Measures

Properly using measurements is the next critical step in empirical analysis after data collection. It is the crucial stage transforming abstract theoretical concepts to operable data for later empirical tests, which directly determines the internal validity of the study. According to Hunter (2001) and Hamermesh (2007), internal validity can be divided into two types, statistical replication and scientific replication. Besides the latter about the external validity, the former particularly relates to the choice and manipulation of measurements, which refers to that "[w]hen a researcher uses a different sample from the same population to evaluate the same theoretical implication as in the previous study with equivalent construct validity..." (Morton and Williams 2010, 258). Stated differently, to validate a study, the researcher should conduct tests on theoretically equivalent measures all the way. This criterion should be taken into account not only in the robustness tests section of a study but also when the main variables are originally measured.

To achieve this criterion, a researcher should consider about three "consistencies": first, *data source consistency.* All the steps in designs and implements of data collection may lead to substantively different results. To minimize measurement error or at least keep it in a coherent way so as to eliminate later, we suggest researchers to use consistent data source, in which the data are collected and coded in a consistent manner. This becomes an issue especially if the researchers is considering create measures from multiple available data sources . In such case, except for validating the measure in general fit the theory they intend to test, the researchers should carefully review the data collection and coding process to ensure that the data collected in each data set reflect the same concept. For example, in a cross-survey study, one should select data produced by the similar questions for each variable.

Taking the aforementioned case, Newman, Johnston, and Lown (2015*a*), for measuring the dependent variable, they used 2005, 2006, 2007, and 2009 surveys on the US citizens conducted by Pew Research Center. For 2005 and 2006, they used responses from exactly the same question. This strategy guarantees data coming from the same organizers, for the same question in same area, and therefore reduces the risk to violate the inconsistency in the data. Researchers should also be aware that any above "sames" singly is not a sufficient condition for measurement consistency. In the previous case, for data from 2007 to 2009, researcher still chose the data from the same organizer and in the same area, but with different questions and coding methods. As shown in the later test, the measurement no longer captured the same concept. Another example is from Asia Barometer. Chu and Huang (2010) found that responses on the same question about the attitude towards democracy did not capture the public attitude to democracy in the theoretical sense, but reflected different understandings of this concept in the given political environment (see, also, Lu and Shi 2014).

Unfortunately, in reality, we may not always enable to keep exactly data source consistency because of the limitation in data availability. In such case, researchers should not let this issue to stop them to sufficiently use all available data (as we suggested in Section 3), while they have to be very careful to combine and manipulate data. Doing them inappropriately may distort data and introduce serious biases into the analysis. To minimize this risk, researchers are ought to consider at least two issues: *structural consistency* and *method consistency.* The structural consistency refers to the consistency in the data coding format. Different format may capture different aspects of a concept which may not be comparable. A famous case of it is the measurement debate on challenges of Przeworski et al. (2000) on the classic developmentalist argument about democratization by Lipset (1960). The former went against the conclusion of the latter arguing that economic development is not a necessary condition for democratization. Nevertheless, in the empirical tests, Przeworski et al. used a binary variable for degree of democracy and only used income to measure development. This led the concern about whether their test truly reject Lipset's theory or just capture a different aspect of the democratization process (not very sure if this citation is adequate Bernhagen 2009). Another example is from the previous case of Newman, Johnston, and Lown (2015*a*).

For 2005 and 2006, they used public attitudes towards two statements "Most people who want to get ahead can make it if they're willing to work hard" or "Hard work and determination are no guarantee of success for most people." If the respondents agree with the first one, they got 1; if they agree with the second, then 0. The result measure is an aggregated binary variable based on two binary survey questions. For 2007, they chose the questions asking the attitudes towards the two statements separately. For each statement, there is a 4-scale categorical recorded from "completely agreed" to "completely disagreed." When respondents answered completely or mostly agree to *both* statements, they got 1, otherwise, 0. For the 2009 data, Newman et al. only used the records for the second attitude, even if the question about the first attitude was also asked in the survey as in 2007. For this year, then, respondents only needed to answer "completely agreed" or "mostly agreed" for the second statement to get 1 in Newman et al. measure of the dependent variable.

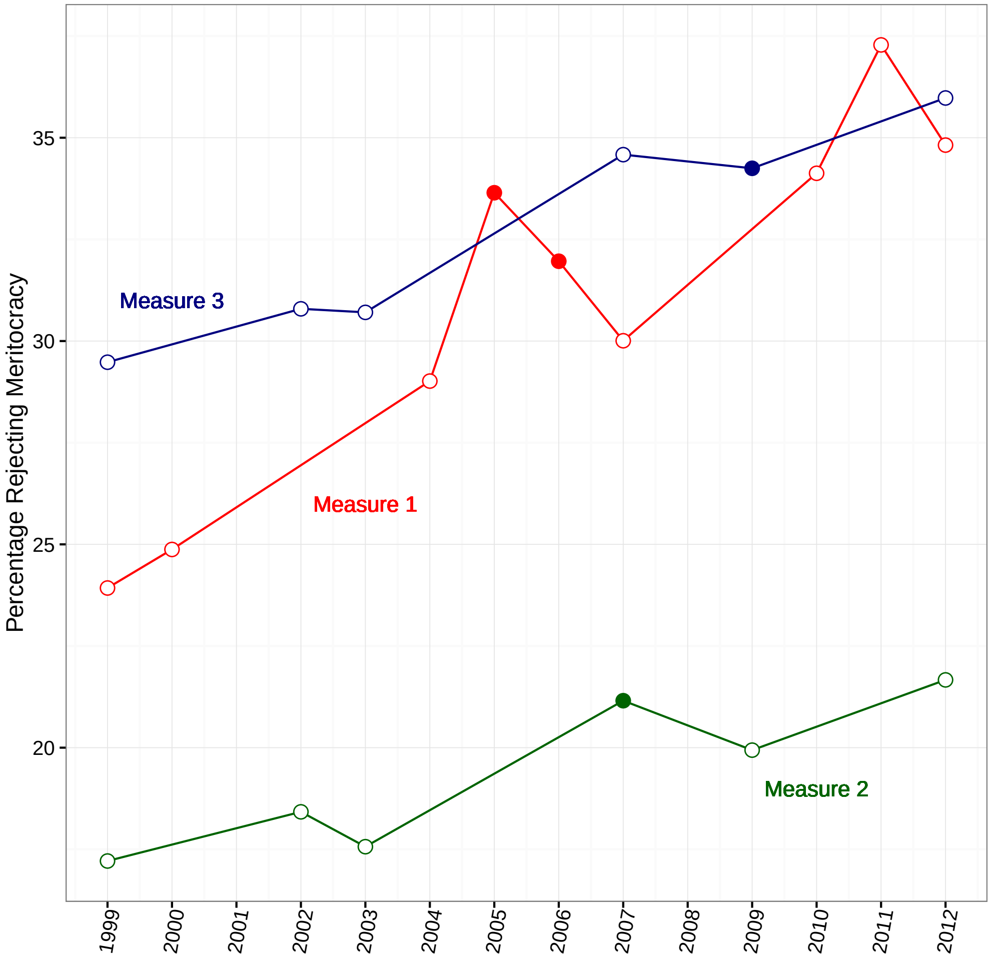
The authors in this case not only used data with inconsistent format, which we admitted being unavoidable only when data with the same structure is unavailable, but also had a method inconsistency issue. They used three different methods to convert survey records to binary. This could make the measure very tricky, since the variance in the measure could also be caused or changed due to the applications of multiple methods.

Our test confirms this concern. We apply each of the three measurements Newman et al. used on structurally consistent data from 1999 to 2012. Figure 3 presents the results. The plot clearly illustrates how different the result from the three methods. Measure 2 is systematically lower than Measure 1 and 3; even if for 1 and 3, they follow very different patterns, except for a couple of dyadic crosses.[3] In this case, a variable consisting of components from these three measures essentially mixed three different things together. No one could exactly know what concept is really measured. In other words, the internal validity has been largely diminished.

To wrap up, in this section, we provide three principles for researchers to protect the internal validity of their studies from scientific malpractice. They are data source consistency, structural consistency, and method consistency. We are fully aware of the tough side in the social scientific studies and, thus, the first two consistencies may not always be achievable. But we also show what risk may be caused if even the method consistency is not held.

[3] We even extend the testing to all the available datasets to 1987 - 2003, and account for the uncertainty in the samples. The result shown in Appendix A manifests that except for a couple of overlaps between two types of measure, most of them are significantly different; nor are cases the three measures towards the same.

Figure 3: Comparing Three Measures of Rejection of Meritocracy Pooled by Newman, Johnston, and Lown (2015*a*)



Notes: The analyses presented in Table 1 of Newman, Johnston, and Lown (2015*a*, 333) were conducted on pooled observations with the dependent variable, rejection of meritocracy, measured in one of three different ways (see Newman, Johnston, and Lown 2015*a*, 331). Here, solid circles represent the data used by Newman, Johnston, and Lown (2015*a*); hollow circles represent data in other available Pew surveys. Plotting the percentage of (weighted) respondents to reject meritocracy by each of these measures reveals that the second measure results in much lower levels of rejection of meritocracy than either of the others and the third often yields considerably higher levels than the first. In light of the evident lack of comparability of these three measures, pooling them into a single analysis cannot be justified.

5 Wrangle Data with Care

Bigger, and better. In survey studies, When conducting quantitative analysis, Scholars look forwards to have a sample with large-N to test more variables and their relationships in statistical models. However, it is difficult to ensure the quality of all surveys in reality: they can be underrepresentative of some specific populations because of non-responses or biased questionarie design, or the sample size is too small due to the resource limit of survey institutes. In order to solve this problem, scholars usually combine different surveys projects or several waves of the same survey together to accquire a satisfising size of sample.

Although it is a common practice, there are several important issues should be noted here. In general, after merging multiple data, scholars should scrutinize whether the numbers of the same variable in different data are identical or not. Second, scholars should be always cautious about the consistency and accuracy of the variable measured in different surveys. This requires cautions to both the questions and the measures of variables.In the ideal situation, the format, wording and measure of questions with regards to the specific variable of interests should be the same across different surveys. If scholars are comfortable with questions of the specific variable are asked differently, they will still need to defense the combination of these questions with evidence that they are actually asking the same thing and will not differentiate survey takers' opinion dramatically. Many empirical studies show that even minor changes in question languages can vary respondents'answer becasue of their different cognition and reception of the information. (**?**)(Rasinski 1989)(Zaller 1992)(Bertrand and Mullainathan 2001) For example, changing a statement from "climate change" to "global warming", respondents' agreement sharply drops by 26.2%. (Schuldt, Konrath, and Schwarz 2011) Furthermore, schoalrs should also need to be very careful in maintaining the measurement consistency in coding and recoding processes. This means scholars should not only unify the measures for the specific variables of interests in different surveys, but also ensure the measures they choose to adopt are accordant with the common practices in the field to ensure comparability with other studies, unless there are specific reasons to use a different measure.

These issues can be easier said than done. The example article used in this paper clearly show that even experienced scholars can fall on these simple issues. First of all, some data do not match with each other after merging the four surveys. When we draw the data out from the countries analyzed in Table 1 and Table 2 (Newman, Johnston, and Lown 2015*a*), it is obvious that the Bush share of the vote in the 2004 election are not exactly match with each other: among all the counties examined in both tables, only fewer than 10% have matching data (even when rounded to two decimal places). We double check the data with the source listed by the authors (http://uselectionatlast.org), and the Bush share of vote in Table 1 corresponds with it. Thus, the data in Table 2 is inaccurate.

Table 1: Mismatched Data on Bush Vote, from Replication Data

County	Table 1 Data	Table 2 Data
Baldwin, AL	0.76	0.79
Calhoun, AL	0.66	0.66
Chambers, AL	0.58	0.56
Cherokee, AL	0.65	0.65
Choctaw, AL	0.54	0.51
Clarke, AL	0.59	0.57

Notes: Newman, Johnston, and Lown (2015*b*) replication data on the share of the vote won by Bush in the 2004 presidential election. The first six counties, when listed alphabetically by state and county, are shown; they reveal that the data employed in Table 2 only occasionally matches that employed in Table 1, even when rounded to two digits. Overall, these data match for fewer than 10% of all counties.

Second, the authors have some unreasonable practices in their paper with regards to variable measures in the coding process. For example, they use a 5-point scale measure for respondent's party identification instead of the the standard 7-point scale party identification is widely applied in classic American politics. According to this measure, 5 and 1 represent strong partisanship , 4 and 2 represent weak partisanship, and 3 represents independent party identification. However, the questions asked in the Pew survey reveals more about the party inclination of indpendent voters. The survey firstly asks whether respondents identify their own party, and how strong such identification is. If respondents identify themselves as indpendent, the survey continue to ask which party they lean to. These questions provide more information about independent voters' party incliation, and we cannot find any reason to throw away this information in this study by adopting the 5-point scale partisanship measure. Besides, the measure of unemployment used in the paper is also problematic. In the 2006 Pew Immigration Poll, unemployment is measured by two questions: (1) whether you are working (either full-time or part-time employed) or not, and (2) if you are not working, which is your current status (student, homeaker, retired, or unemployed). However, in the

2005 Pew News Interest Index Poll, 2007 and 2009 Pew Values Survey, only the first question was asked, and the authors treat respondents in the combined dataset who are not working as ”unemployed”, ignoring the detailed employment status difference of the 2,000 respondents of the 2006 survey.

These malpractices seems small, but they may dramatically change the result of statistical model when inputting wrong data and severely undermine the validity of the study. To prevent these problems, every scholar to be more scrutinized in merging data. They need to review all the questions and variables before merging, reassure the measures of variables across dataset are consistent and valid in the recoding process, and they need to be more transparent in data coding.

6 Multiply Impute Missing Data

Summary Paragraph

In political science study, a common way to contaminate the data and undermine the validity is missing data growing non-response in Pew survey (Curtain, et.al, 2005;);

Reason: poor survey design(literature needed); private information (income;); sensitive question—influence of social desiablity (underreported abortion of black: Jagannathan, 2001; nonrespond and overreponse in female president question; Streb, 2008; nonresponse in religious: Hadway et.al, 1993); survey forms (Chang RDDv.s. Internet; 2009; Amazon-Turk, Berinsky, 2012)

To deal with missing data: first identify missing mechanisms: (1) missing completely at random (MCAR): the probability of missingness is the same for all unit; that is, each survey respondent choose not to answer the question based on on coin flip (2)missing at random (MAR), and nonignorable or (King et. al, 2001): the probability whether the survey question is answered may depend on the other factor which are observable in the data: for example, an independents are more tend to decline to answer partisan identification question. (3)nonignorable (NI): because of the unobserved value of the missing response: high income people tend to conceal their real income, and other variables in the data cannot predict which respondent have high income.

Strategy: to all: 1. listwise deletion or complete-case analysis; + Available case study (use the distribution of other observable variables)

Disadvantages: (1) lead to biased estimates, especially when missing values differ systematically from the completed observed cases; (2) relatively, the standard error can be sensitive due to the original sample size and the deleted case size; (3) in ACS, may lead to omission of a variable hat is necessary to satisfying the assumptions necessary for desired casual inference.

2. Non-response weighting: need to add literatures (for MAR and MCAR)

3. Simple imputation: With high certainty: (1)mean, (2)last value carried forward;

With uncertainty: (3) using information from related observation (for example, in GSS, using reported occupation types and its mean annual salary to infer income; or in SIS, use reported working months to infer income)

Disadvantages: (1) mean, last value, or other singly imputed method bias, especially when the size of missing observation is relatively large; distorting the actual distribution (Gelman, 2006)

(2) Inferred from other related observation: may not to impute all missing data

4. Random Imputation with a single variable or multiple variables

5. Multiple Imputation:

Definition: imputing missing values for each missing case with different imputations to reflect uncertainty levels, and creating a complete data set. (King.et.al, 2001). For example, if assume the missing data of a specific variable is MAR, indicating other observable variables can infer useful information to predict the missing cases, conditional on whichever imputation model adopted.

(1)Multivariate regression:

a. use continuous model to impute missing discrete response.Example: modeling the data as continuous (transfer discrete value into continuous), and imputing continuous values (Gelman and King, 1998)

(2)

Specific to Newman's paper

church attendance—all missing are simply assigned "once or twice a month"

income, a variable of interest, is missing for over 10% of the sample, but values are mysteriously single-imputed (where did these values come from? they aren't meaningful–they fall between categories)

ideology, partyid also single-imputed, it seems

Church attendance: MAR

income: MAR

ideology: MAR
Thus, mulitple imputation is possible and necessary
missing data should be multiply imputed (e.g., King et al. 2001)

7 Plot Interaction Terms

It has been well known for over a decade that models containing multiplicative interaction terms require particular care in interpretation (see, e.g., Golder 2003; Braumoeller 2004; Brambor, Clark, and Golder 2006; Kam and Franzese 2007). Nevertheless, many political scientists struggle with interaction terms: improperly specified or interpreted interaction terms appear at the top of Nyhan's (**?**) list of "recurring statistical errors" that reviewers should be sure to check for.

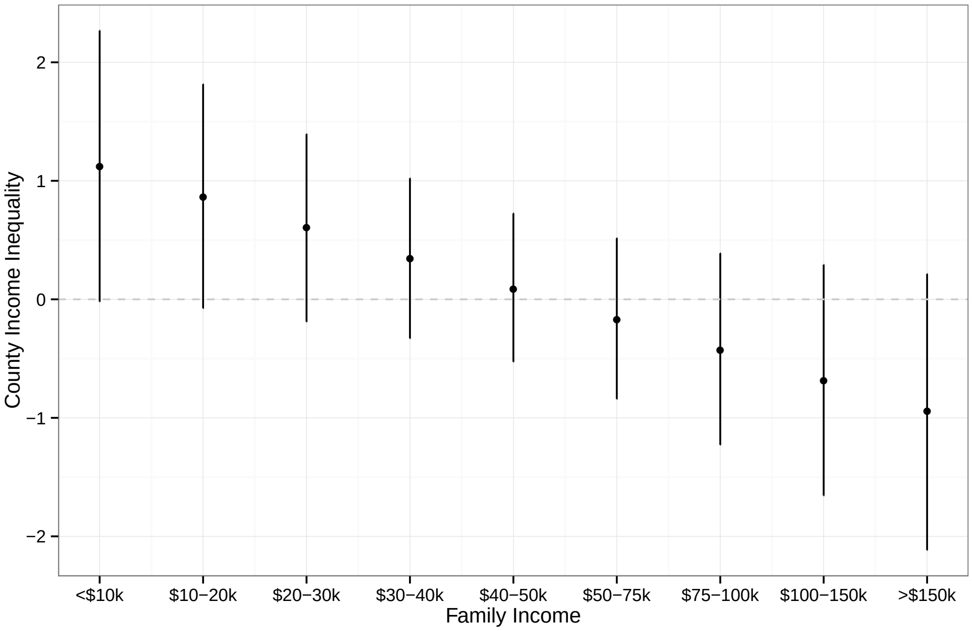
First, as Brambor, Clark, and Golder (2006, 71-72) wrote, in models with multiplicative interaction terms, constitutive terms should not be interpreted as unconditional or marginal effects because "the coefficient on X only captures the effect of X on when Z is zero."

Given this, in Table 1, the coefficient on "Median Household Income (X)" only captures the effect of "Median Household Income (X)" on "Rejection of Meritocracy (Y)" when "GINI Index (Z)" is zero. Similarly, the coefficient on "GINI Index (Z)" captures the effect of "GINI Index (Z)" on "Rejection of Meritocracy (Y)" when "Median Household Income (X)" is zero. In this sense, we cannot conclude that "the estimates (of GINI Index) reveal that, among those with the lowest incomes, an increase in county inequality is associated with a significant increase in the probability of rejecting meritocracy" (Newman, Johnston, and Lown 2015*a*, 334) because the lowest value of "Median Household Income" variable is not 0, but 1 (Newman, Johnston, and Lown 2015*a*, 332).

Second, more importantly, the statistical significance of the interaction term does not mean that X has substantive conditional effect on Y. The table itself does not give us the sufficient information because the marginal effect of X on Y should be interpreted with the substantively relevant values of the variable Z. Moreover, as Brambor, Clark, and Golder (2006, 74) demonstrated, we cannot calculate the standard errors from the typical results table using a little algebra. Thus, a figure as Figure 1 is needed to show the marginal effect of X and the corresponding standard errors. Figure 1 plots the coefficient estimates for county income inequality (GINI Index) at each of the nine levels of income in the Pew data. In spite of the statistical significance of the interaction term in Table 1 of the original paper, the confidence intervals of these estimates all cross zero. In other words, none are statistically significant. The conclusion of Newman, Johnston, and Lown (2015*b*, 334) that this result "reveals that among low-income citizens, those residing in highly unequal contexts are significantly more likely to reject meritocratic ideals than those in relatively equal contexts" is therefore erroneous.

Actually, Newman, Johnston, and Lown (2015*b*, 334) did the similar work to show the substantive conditional effect as Figure 2 of their paper. However, this graph is not only insufficient to show the substantive conditional effect, but also it is not reproducible. They indicate that Figure 2 "plots the predicted probability of rejecting meritocracy across levels of county inequality for citizens at the 5th and 95th percentiles of household income", but we could not obtain the same results using the variable "Median Household Income" that ranges from 1 to 9. This is because in their published replication data for the Model 1 (White Respondents) of Table 1, the range of the variable "Median Household Income" is oddly .21 to 1. In sum, the conditional effect is statistically significant only when the variable "Median Household Income" ranges from .21 to 1, which is substantively meaningless since there are no respondents with values of "Median Household Income" below 1.

Figure 4: Logit Coefficients of Local Income Inequality by Respondent Income: Table 1, Model 1, From Replication Data



Notes: The coefficient for county income inequality fails to reach statistical significance for any observed level of respondent family income.

References

Appadurai, Arjun. 2000. "Grassroots Globalization and the Research Imagination." *Public culture* 12(1):1–19.

Asendorpf, Jens B, Mark Conner, Filip De Fruyt, Jan De Houwer, Jaap JA Denissen, Klaus Fiedler, Susann Fiedler, David C Funder, Reinhold Kliegl, Brian A Nosek et al. 2013. "Recommendations for Increasing Replicability in Psychology." *European Journal of Personality* 27(2):108–119.

Bernhagen, Patrick. 2009. "Measuring Democracy and Democratization." *Democratization* pp. 24–40.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean

What They Say? Implications for Subjective Survey Data." *American Economic Review* pp. 67–72.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82.

Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58(4):807–820.

Brunswik, Egon. 1955. "Representative Design and Probabilistic Theory in a Functional Psychology." *Psychological Review* 62(3):193–217.

Chang, Andrew C, Phillip Li et al. 2015. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not". Technical report Board of Governors of the Federal Reserve System.

Chu, Yun-han, and Min-hua Huang. 2010. "The Meanings of Democracy: Solving an Asian Puzzle." *Journal of Democracy* 21(4):114–22.

Collaboration, Open Science. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251).

Dafoe, Allan. 2014. "Science Deserves Better: The Imperative to Share Complete Replication Files." *PS: Political Science & Politics* 47(1):60–66.

Denzin, Norman K, and Yvonna S Lincoln. 2009. "Qualitative Research." *Yogyakarta: PustakaPelajar* .

Golder, Matt. 2003. "Electoral Institutions, Unemployment, and Extreme Right Parties: A Correction." *British Journal of Political Science* 33(3):525–534.

Hamermesh, Daniel S. 2007. "Viewpoint: Preplication in Economics." *Cadnadian Jounal of Economics* 40(3):339–353.

Hunter, John E. 2001. "The Desperate Need for Replications." *Journal of Consumer Research* 28(1):149–158.

Jacoby, William G. 2015. "The AJPS Replication Policy: Innovations and Revisions.".

Janz, Nicole. 2015. "Bringing the Gold Standard into the Classroom: Replication in University Teaching." *International Studies Perspectives* pp. 1–16.

Kam, Cindy D., and Robert J. Franzese. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28:444–452.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Lipset, Seymour Martin. 1960. *Political Man: The Social Bases of Politics*. Baltimore: Johns Hopkins University Press.

Lu, Jie, and Tianjian Shi. 2014. "The Battle of Ideas and Discourses before Democratic Transition: Different Democratic Conceptions in Authoritarian China." *International Political Science Review* p. 0192512114551304.

Lupia, Arthur, and Colin Elman. 2010. "Memorandum on Increasing Data Access and Research Transparency (DA-RT)." *Submitted to the Council of the American Political Science Association, September* .

Lupia, Arthur, and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47(1):19–42.

Morton, Rebecca B, and Kenneth C Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

Newman, Benjamin J., Christopher D. Johnston, and Patrick L. Lown. 2015*a*. "False Consciousness or Class Awareness? Local Income Inequality, Personal Economic Position, and Belief in American Meritocracy." *American Journal of Political Science* 59(2):326–340.

Newman, Benjamin J., Christopher D. Johnston, and Patrick L. Lown. 2015*b*. "Replication data for: False Consciousness or Class Awareness? Local Income Inequality, Personal Economic Position, and Belief in American Meritocracy." http://dx.doi.org/10.7910/DVN/26584, Harvard Dataverse, V2.

Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990.* Vol. 3 Cambridge University Press.

Rasinski, Kenneth A. 1989. "The Effect of Question Wording on Public Support for Government Spending." *Public Opinion Quarterly* 53(3):388–394.

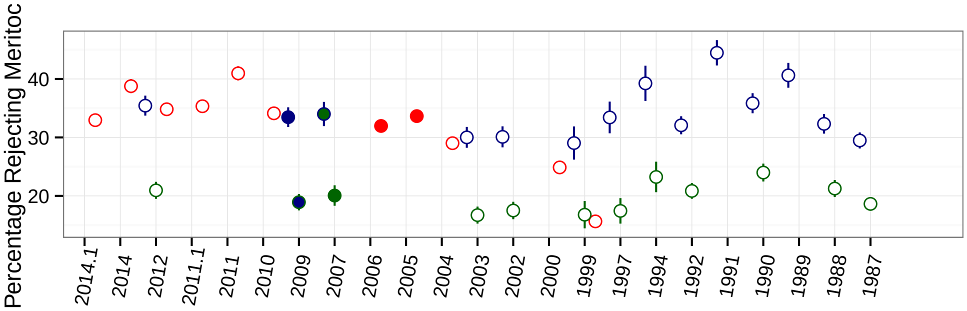
Schuldt, Jonathon P, Sara H Konrath, and Norbert Schwarz. 2011. ""Global Warming" or "Climate Change"? Whether the Planet Is Warming Depends on Question Wording." *Public Opinion Quarterly* p. nfq073.

Young, Joseph K, and Nicole Janz. 2015. "What Social Science Can Learn From the LaCour Scandal." *Chronicle of Higher Education* .

Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* Cambridge University Press.

Appendices

A Measurement Comparison with Uncertainty



Notes: The graph replicate what we did in Figure 3 but with all available data and uncertainty estimations.