

SECOND DECLARATION OF DR. ALAN J. SALZBERG

Dr. Alan Salzberg, for his second declaration pursuant to 28 U.S.C. § 1746, deposes and says as follows:

I. Introduction

1. I am the Principal (and owner) of Salt Hill Statistical Consulting. I previously submitted a declaration in this case, dated February 14, 2019 (“Salzberg February Declaration”). My February declaration commented on the “Declaration of Jonathon Penney” (“Penney December Declaration”), which was submitted in December 2018. I submit this second declaration at the request of the United States Department of Justice in response to the “Reply Declaration of Jonathon Penney” (“Penney Reply”), which was submitted on March 8, 2019. I have previously submitted my resume describing my background and qualifications in statistical sampling, analysis, and review for government and industry, as well as information regarding prior testimony and fees.
2. This report proceeds as follows. In the next section, I summarize my findings. In the third section, I detail those findings. In the fourth section, I set forth my conclusions. Finally, I have included an appendix with a program log showing the results of additional analyses.

II. Summary of Findings

3. In my February Declaration I addressed the deeply flawed model presented by Dr. Jonathon Penney in his December Declaration. Specifically, in summary I previously found as follows:¹
 - A. “The methodology used in the Penney Declaration—which purportedly shows an upward trend in page views of certain articles posted on Wikipedia through May 2013, followed by an abrupt drop and downward trend in views of those articles beginning in June 2013—is deeply flawed, inappropriate, and likely biased.”
 - B. “The Penney Model simply assumes that a single change occurred in June 2013, rather than letting the data identify the timing and number of changes in trends that occurred. Even though there is no consistent trend in the data, the design of the Penney Model will create the appearance that the data contain just one inflection point. And, because of its design—even though changes in trend occurred before these June 2013 disclosures—the Penney Model will find that the disclosures caused them.”
 - C. “Contrary to the hypothesis presented in the Penney Declaration, analysis of page views for the 48 individual articles in the privacy-sensitive group do not show a rising trend followed by an immediate and sustained drop in June 2013.”
 - D. “With the one exception of removing the article on Hamas, the Penney Declaration does no analysis or adjustment for factors (such as world events) affecting these individual article page views. Instead, the Penney Declaration inappropriately aggregates the vastly different page view data for individual articles, with the result that these individual differences in page views are masked.”

¹ Salzberg February Declaration, paragraph 4.

- E. “Even at that aggregate level, I find that the hypothesized peak in page views of “privacy-sensitive” articles in May 2013 does not exist, and the hypothesized upward and then downward trends in views of privacy-sensitive articles before and after June 2013, respectively, do not exist.”
 - F. “Extended data through 2018 regarding page views of the privacy-sensitive articles do not indicate a long-term decline in page views from pre-June 2013 levels.”
 - G. “A proper control dataset would exhibit similar page view behavior prior to June 2013. The comparison datasets used in the Penney Declaration do not and are thus inappropriate controls.”
 - H. “The Penney Declaration analysis ends in July 2014. No data are presented that shed any light on whether page views at the time the Amended Complaint was filed in 2015 (or thereafter) were affected by Upstream. In other words, even if the purported effect and trends were a correct conclusion for the data examined (and they are not), the Penney Declaration analysis does not and cannot show that the effect continued years after the study ended.”
 - I. “Even if a chilling effect occurred in June 2013, there are no data analyzed in the Penney Declaration that show any effect was due specifically to “public awareness of” the specific NSA surveillance program challenged here (known as Upstream surveillance) rather than possible inaccuracies, if any, about the program reported in the press, disclosures about other NSA programs, disclosures about other surveillance programs (e.g., surveillance by Britain), or other, unrelated events of June 2013.”
4. As discussed in detail below, the Penney Reply does not raise any valid critiques of my original findings, and the additional analyses in the Penney Reply do not bolster the flawed model presented in the Penney December Declaration. In addition, the Penney Reply does not propose a new model that corrects the flawed model presented in the Penny December Declaration, and the slight modifications attempted do not address any of the issues I raised. Therefore, my findings and conclusions set forth in my February declaration remain unchanged.

III. Details of Findings

5. The Penney Reply begins with seven critiques of my analyses, in paragraphs 4 through 23 of the Penney Reply, and goes on to respond to my critiques in paragraphs 25-36. I reviewed all of the Penney Reply and in this declaration I organize my responses by topic, so as not to be repetitive. In particular, this section proceeds with the following six subsections:
 - A. Overview of the Incorrect Assumptions Made in the Penney Reply;
 - B. Spurious Statistical Conclusions from the Penney Model are Partly Due to Aggregation of the Article View Data;
 - C. The Penney Reply’s Additional Analyses Fail to Address the Flaws in the Penney Model;
 - D. Data Beyond Time Period 2014 Show Article Views at About 2012 through 2014 Levels, Even When Earlier Data is Corrected for Mobile Views;;
 - E. Omitted Variable Bias of the Penney Model Cannot be Solved by Deleting Valid Data; and

F. The Penney Model's Failure to Isolate the Effect of Awareness of NSA's Upstream Program.

A. Overview of the Incorrect Assumptions Made in the Penney Reply.

6. Before responding to the specific claims of the Penney Reply, I first address some false assumptions the Penney Reply made regarding my critiques.
7. First, while my review of the disaggregated data provides an important, simplified explanation of many of the flaws of the Penney Model, the flaws of that model remain whether that model is applied to the aggregated or the disaggregated data. I am not suggesting that the application of the deeply flawed Penney Model to each of the 48 articles, individually, would be appropriate, nor am I suggesting that there could never be theoretical circumstances where the data could be aggregated without presenting the deeply flawed and misleading results that the Penney Model presented here.
8. Second, my February report provided no wholesale critique of the so-called ITS "Interrupted Time Series" designs or of regression models in general. My critiques instead relate to the particular methods Dr. Penney employed and the underlying data used in the Penney December Declaration.
9. Third, as I pointed out a number of times in my February Report,² I do not present an alternate model of page views, but I do use a number of examples and perform analyses that demonstrate the flaws in the Penney Model. Statements in the Penney Reply regarding "alternatives" that I suggest are therefore misleading.
10. Fourth, while the Penney Reply is correct in that much of my analysis uses "visual inspection" as an aid to understanding the issues with the Penney Model, I also perform statistical tests and point out many specific flaws in the Penney December models. As with the issue of aggregation and disaggregation, I am not advocating one or the other, but rather, doing both. A simple visual review of the data using charts and graphs, such as I the one performed, has long been considered a fundamental component to developing correct statistical models.

B. Spurious Statistical Conclusions from the Penney Model are Partly Due to Aggregation of the Article View Data

11. As I stated in my original declaration, a review of the disaggregated data leads to the conclusion that there is no May 2013 peak or steep drop beginning in June 2013, contrary to the Penney December Declaration's conclusion.³ The aggregated data do not show a May 2013 peak either, but rather an April 2013 peak,⁴ followed by declines beginning in May 2013. This means that both the disaggregated data and the aggregated data show that the drop in page views begins *before* the June 2013 disclosures.

² In my February Declaration, for example, in paragraphs 52 (note 31) and 53, I explicitly state that I am not proposing an alternative model.

³ Salzberg February Declaration, paragraphs 11-17.

⁴ As stated, while the Boston Marathon bombings are one possible reason for an April peak, such a simple model is far from adequate for many of the same reasons that the Penney Model is inadequate, but at least in the April peak model, the drop in page views comes after the purported cause and not before the purported cause.

12. The Penney Reply responds to this critique by erroneously claiming that my disaggregated review should be disregarded because aggregation is appropriate here. This somewhat misses the point, which is that neither the disaggregated data nor the aggregated data support Dr. Penney's conclusions. Nonetheless, I reviewed the literature that the Penney Reply cites in support of its claim. The literature cited does not support the Penney Reply's claim.⁵ The articles cited are general articles on ITS designs rather than articles espousing aggregation.⁶ Moreover, the four reasons cited in the Penney Reply paragraph four are not specific to aggregated data (they apply equally to disaggregated data).
13. In paragraph 26b, the Penney Reply acknowledges that there is "there is no single determinative method or factor to decide whether an aggregated or disaggregated analysis of data is appropriate." As I stated, by aggregating all the data prior to analysis, there is no possibility of correcting for any article-specific differences in the data or exploring whether there are important differences among article views that need to be accounted for in any model. This leads to a biased model and erroneous claims of statistical significance where, as here, such differences exist.⁷ The idea of not reviewing and understanding the disaggregated data, and discarding information by inappropriately aggregating that data, is anathema to scientists, because such ignorance often leads to false conclusions. That review typically includes graphical analysis, because, as one statistician put it: "[g]raphics reveal data. Indeed graphics can be more precise and revealing than conventional statistical computations."⁸
14. The Penney Reply argues that that my use of simple graphs to provide a visual inspection of the disaggregated data should be disregarded, in part, because a "visual inspection of data . . . can often be misleading," a point he makes with a quotation of one of the great proponents of graphical analysis, Dr. Howard Wainer.⁹ Dr. Wainer, however, is not saying that graphs should not be used; he is only saying to be careful that they are not used in a misleading manner.¹⁰ Ironically, by ignoring the disaggregated data and aggregating dissimilar page views to tell a

⁵ Specifically, in the footnotes for paragraphs 4-5, the Penney Reply identifies several sources that Dr. Penney claims supports his use of aggregation in this circumstance. The only citation that even appears to support aggregation, in this type of situation, is not from a paper or textbook but from a PowerPoint presentation by Emma Beard which appears to have been presented at a conference in London (see footnotes 3, 4, and 21 in the Penney Reply). I reviewed the PowerPoint presentation and it presents no reasoning or data to support the claims (nor is it obvious that the author even made such claims regarding a simple regression model like the one in the Penney December Declaration). Additionally, unlike a scholarly article, a PowerPoint presented at a conference is typically accompanied by an oral portion of the presentation that may provide additional context or present the point differently than the language on the printed slides). In short, none of the cited source materials in the Penney Reply alter my conclusion that in this instance the use of aggregated data is inappropriate and misleading.

⁶ The Penney Reply in paragraph 26e, takes issue with my terming the data "panel data" and not "time series" data. Panel data is a form of time series data, as the introduction to the text and chapter on panel data in my source make clear. See Wooldridge, Jeffrey M., Introductory Econometrics, A Modern Approach, 5th Edition, 2012, South-Western Cengage Learning, p. 10 and 448.

⁷ See Salzberg February Report, paragraphs 56-60, for example.

⁸ Tufte, Edward R., The Visual Display of Quantitative Information, Graphics Press LLC, 2001, p. 13. Also, p. 9 of the same text states that: "Often the most effective way to describe, explore, and summarize a set of numbers – even a very large set – is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful."

⁹ Penney Reply, paragraph 3 footnote 1.

¹⁰ Quote from Penney Reply, paragraph 3 footnote 1 (quoting Howard Wainer).

misleading story, Dr. Penney has created precisely the type of misleading graphs that Dr. Wainer is warning against. As Yale statistician Edward Tufte says: “[a]ggregations by area can sometimes mask and even distort the true story of the data”¹¹ and “[a]ggregations over time may also mask relevant detail and generate misleading signals.”¹² Tufte concludes: “[i]f in doubt, graph the detailed underlying data to assess the effects of aggregation.”¹³ As I explained in my first Declaration (Paragraphs 18-26), Figure 2 of the Penney December Declaration is misleading because it inappropriately aggregates the data and shows a suggestive regression line, while obscuring the fact that the decline was not as indicated.¹⁴

15. I created and included (as Appendix IV to my first Declaration) graphs of each of the 48 articles’ page views individually, so that all the data is available to view in a clear graphical form. I invited (and invite) review of each of those graphs. The only reasonable conclusion from a review of those graphs is that the effect supposedly found in the Penney December Declaration is spurious. I also included graphs of the aggregated data (*see* paragraphs 18-26 of my February Declaration), and those graphs also do not indicate a May 2013 peak. I did not leave out anything or “cherry-pick,” contrary to what the Penney Reply states in paragraphs 11, 12, and 32(a).¹⁵
16. The Penney Reply claims that disaggregation adds “noise” to the data, “both visual and statistical,” and points to my first graph showing all 48 articles in a single figure.¹⁶ I showed all the data in a single figure (as well as in 48 separate figures in Appendix IV) because it provides important context and a comparison point to Dr. Penney’s aggregated plot, which artificially smooths the differences.¹⁷ As a reminder, my Figure that includes all 48 articles is below.

¹¹ Tufte, Edward R., *Visual Explanations*, Graphics Press LLC, 1997, p. 35.

¹² *Ibid*, p. 36.

¹³ *Ibid*, p. 37.

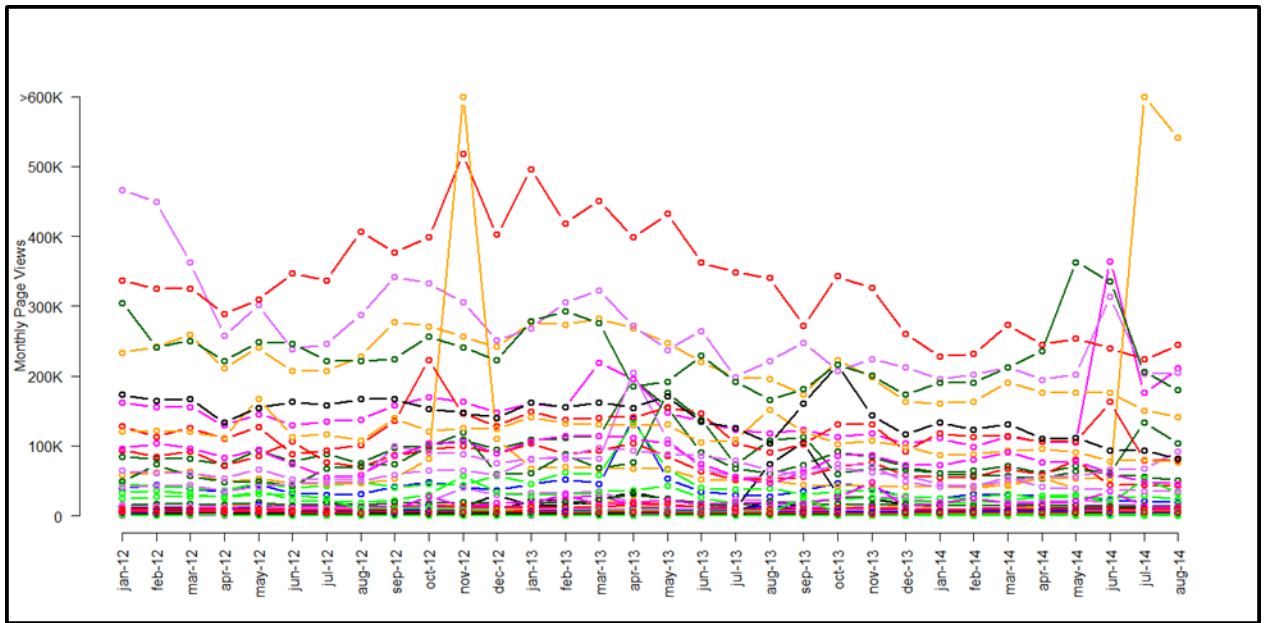
¹⁴ Gelman, Andrew and Zelizer, Adam, “Evidence on the deleterious impact of sustained use of polynomial regression on causal inference,” *Research and Politics*, January-March 2015, also cited in the Penney Reply, is also clear that graphical analysis is recommended.

¹⁵ Penney Reply, paragraph 12.

¹⁶ Penney Reply, paragraph 5.

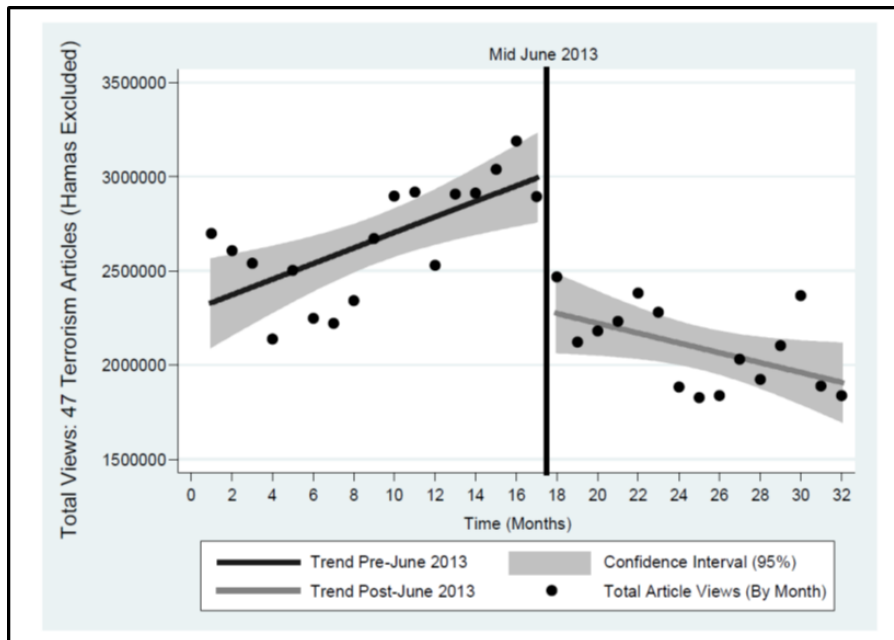
¹⁷ This method, of putting all the data into a single plot, is done in so-called spark graphs, examples of which can be found in Tufte, Edward R., *Beautiful Evidence*, Graphics Press LLC, 2006. p. 47-63.

Figure 1: Page Views for all 48 Articles Considered to Suffer from a Chilling Effect beginning in June 2013



17. In contrast, Penney’s Figure 2 from his December Declaration, shown below, misleadingly indicates a simple up and down movement that is belied by the individual data in Figure 1, above. The same is true for the confidence intervals drawn on Penney’s Figure 2 graph itself, as I pointed out in my February declaration.¹⁸

¹⁸ See Salzberg February Declaration, paragraph 20.

Figure 2: Penney December Declaration Aggregate Figure Masks Individual Differences

18. The contrast between the simple, disaggregated view and Penney’s misleading aggregate view should have led Dr. Penney to question whether his aggregated model masks systematic differences at the article-level. As I stated above, I also provided each plot individually, so the reader can see what is behind the total picture of page-views shown in my first figure.
19. The Penney Reply is also wrong when it categorically states that disaggregation adds noise. The Penney Reply concerns that a disaggregated model will not allow for estimation of an “aggregate level inference about large scale NSA surveillance effects”¹⁹ are misplaced. If the same naive model is run on both datasets, the estimated effect in the disaggregated model is exactly the same as the estimated effect in the aggregate model.²⁰ The statistical significance of these effects will also be the same if the disaggregation only adds noise to the model, and I show this fact through a simulation.²¹
20. However, if the disaggregated data reveal systematic differences in the data, in that the individual articles’ page views do not tell the same or even a similar story as the aggregated data, then the naive model needs to be modified in order to avoid bias, whether run on aggregate or individual article data. To further support the analysis I already performed showing the model is over-simplified and perhaps mis-specified, I performed a statistical test to determine whether the

¹⁹ Penney Reply, paragraph 26b suggests that because the question regards aggregate differences the aggregated data must be used.

²⁰ This fact is shown in the Appendix to this Declaration, and can be observed by noting the coefficient estimates for the Penney Model as shown in my Appendix. In the Appendix, I run the Penney Model on the averages and run the same model on the individual articles. The estimated effects (model coefficients) are exactly the same.

²¹ I have included in the Appendix a simulation that shows the results of running the Penney Model on aggregated and disaggregated data are the same when the errors are statistical noise. This includes not only the regression coefficients (which will be the same whether the difference are due to noise or not, as explained above) but also the standard errors (i.e., the statistical significance) of those coefficients.

differences by article are mere noise or systematic.²² I found, with high statistical significance, that the differences among articles are systematic (the statistical results are in the Appendix to this Declaration). This means that the model used in the Penney Reply is incorrect, *whether using the aggregated or the disaggregated data*. Only by accounting for the article-level, seasonal, and other differences can a valid model or set of models be produced. Furthermore, the model's estimates show increased error when calculated in disaggregated form.²³ This fact confirms my conclusions in my February report.²⁴ Because the differences are systematic and not mere "noise," the aggregation produces a result with inflated statistical significance.²⁵

21. In reviewing some of the specific examples I cited to explain the fact that aggregating the data masks differences in the articles, the Penney Reply re-explains some analyses and runs additional models, but none address the issues I raised.²⁶ The Penney Reply presents Figures 2A and 2B, which purported show an "Increase until June 2013 and then a Sharp Drop-off."²⁷ This labeling is wrong. The increase is only through April, with a drop off in May and a continuation of that drop in June. This fact can be seen in Penney Reply's own Figures 2A, 2B, 3A, and 4 of the Penney Reply. Each shows an April and not a May peak, and a May and not a June start to the drop in page views. As I explain in my first Declaration, the fact that the drop in page views began *before* the June 2013 disclosures does not support Dr. Penney's conclusion that the June 2013 disclosures caused the drop in page views, and violates a basic tenet of causal models (i.e., a cause cannot occur after an effect).²⁸
22. The models using the data in Figures 2, 3, and 4 of the Penney Reply suffer from the same problems as the original model in the Penney December Declaration.²⁹ The Penney Reply

²² See Salzberg February Report, paragraphs 55-60 for my comments regarding the over-simplified model and omitted variable bias.

²³ As shown in my appendix, attached here, in some cases the claimed effects are not statistically significant. In other cases the statistical significance is weaker. These are further indications that the article differences are not mere noise. As my simulation (in the Appendix attached here) shows, when differences are based on mere noise, the statistical significance of the coefficients for the effects will remain unchanged when running the model on aggregated versus disaggregated data.

²⁴ See Salzberg February Report, paragraph 48-50 and 55-60.

²⁵ This is due to omitted variable bias, among other factors. I pointed this out in my February Report, paragraph 56. I do not attempt to correct for the omitted variable bias by adding additional variables, and therefore the disaggregated model is also incorrect.

²⁶ These re-analyses and the Penney Reply's commentary on them is found in Penney Reply, paragraphs 6-22 and paragraphs 26, 28, and 30.

²⁷ Penney Reply, Figures 2A and 2B.

²⁸ For two examples of such spurious inferences that ascribed a later cause to an earlier effect, see a source cited in the Penney Reply: McCleary, Richard, McDowall, David, and Bartos, Bradley J., [Design and Analysis of Time Series Experiments](#), Oxford University Press, 2017. The examples are portrayed in this text in Figure 5.15 (explained on p. 214-215) and Figure 7.1 (explained on p. 275-276), and involve "interventions" and data with similarities to the data analyzed in the Penney December Declaration.

²⁹ The Penney Reply inexplicably discards its high-privacy group of 31 articles in favor of a new high privacy group of 23 articles for Figure 4 and some accompanying analyses. The Penney December Declaration already determined (perhaps also arbitrarily) a 31-article set that is highly privacy sensitive and this new set of 23 is a subset of those articles. Of course, re-running the same model on datasets that are nearly the same will produce results that are nearly the same, and proves nothing.

analysis ignores the large and obvious effect of events of April 2013 in its analysis of “improvised explosive device,” “dirty bomb,” “car bomb,” and “ammonium nitrate.”³⁰

23. The only graph that the Penney Reply shows that appears to have a peak in May is Figure 3b (page views for so-called “normalized” Ammonium Nitrate), but that supposed “peak” is artificially created because the Penney Reply manipulated the graph to remove the April peak and replace it with the average of the March and May.³¹ Removing such outliers and replacing them with averages in this way is against the practice of statisticians in general. Outlier handling is discussed in detail in an article the Penney Reply cites (at footnote 8), and this article says such adjustment is only appropriate for *error* outliers.³² Here, the data points for Ammonium Nitrate page views are not errors and so removing the correct data point and replacing it with an average is inappropriate.³³

C. The Penney Reply’s Additional Analyses Fail to Address the Flaws in the Penney Model

24. Paragraphs 18 and 28 of the Penney Reply assert that no assumption is made in the Penney Model concerning a May peak. However, the Penney Model hypothesis is a single trend line through May 2013, and then a second line, starting in a potentially different place. The assumption is a single point of inflection, and that point is a peak in May and a drop off beginning in June.³⁴ While it is correct that the model can find that there is no peak at all in the data, my point is that no other month is modeled as a possibility, and that if the data goes up and down, the model finding a June peak will be statistically significant even though the peak did not occur in May and the drop did not begin in June.
25. The Penney Reply in paragraph 28 criticizes my demonstration, using a polynomial model, that the peak did not occur in May and says such an approach is biased, citing a scholarly article.³⁵ That article refers to higher order polynomials (which I did not use) and, even for higher order polynomials, the article does not say that such models are biased, only that they may not reduce bias.³⁶ Indeed, as shown in the quote below, the article brings up the same issues that I do with

³⁰ While the Boston Marathon bombings did not use ammonium nitrate and were not a “dirty bomb,” this does not mean they may not have been a reason for a huge uptick in page views. Some news articles (for example <https://www.theatlantic.com/technology/archive/2013/04/new-boston-bomb-parts-photos/316183/>) discussed the possibility of ammonium nitrate being used. Even if the Boston Marathon bombings had nothing to do with the April uptick in page views, the complete exclusion of any cause of those changes biases the Penney Model, as I have explained.

³¹ Penney Reply, paragraph 14 and footnote 8. See page 11 of the Penney Reply for the graph of Ammonium Nitrate views without April data deleted and replaced with the average of March and May 2013.

³² The article is Aguinis, Herman, Gottfredson, Ryan K., and Joo, Harry, “Best-Practice Recommendations for Defining, Identifying, and Handling Outliers,” *Organizational Research Methods*, 16(2), 2013, p. 270-301.

³³ Neither Dr. Penney nor I have suggested that the change in views in ammonium nitrate in April 2013 was due to an error in the archives used to collect the data.

³⁴ Penney December Declaration, paragraph 23, describes the design as testing for a “decrease in level and trend” beginning in June 2013.

³⁵ The article, cited in paragraph 28(b), footnote 32 of the Penney Reply, is Gelman, Andrew and Zelizer, Adam, “Evidence on the deleterious impact of sustained use of polynomial regression on causal inference,” *Research and Politics*, January-March 2015

³⁶ Gelman, Andrew and Zelizer, Adam, “Evidence on the deleterious impact of sustained use of polynomial regression on causal inference,” *Research and Politics*, January-March 2015, p. 5.

respect to simplistic linear models, saying that modeling higher polynomial effects does not necessarily fix those issues:

“the higher-order polynomial has the effect of slightly modifying and improving the fit of the natural linear model. In criticizing the use of high-degree polynomials in RD [RD stands for Regression Discontinuity—the issue theorized in the Penney December Declaration] adjustments, we are not recommending global linear adjustments as an alternative...We recommend that any RD analysis include a plot such as Figure 1 showing data and the fitted model, and that users be wary of any resulting inferences based on fits that don’t make substantive sense.”³⁷

26. In other words, plotting the data is recommended, and the authors are not recommending that a simple linear model is better than a polynomial one. Indeed, they preface that discussion specifically with:

“Our point here is not to argue that the linear model is correct...Our point is rather that the headline claim, and its statistical significance, is highly dependent on a model choice that may have a data-analytic purpose, but which has no particular scientific basis. Figure 1 indicates to us that neither the linear nor the cubic nor any other polynomial model is appropriate here. Instead, there are other variables not included in the model which distinguish the circles in the graph.”³⁸

27. I include these extended quotes because despite the Penney Reply’s misinterpretation, the article is useful in that it points out the very issue of spurious statistical significance and omitted variable bias that is at the heart of my critiques of the Penney Model in the first place.

28. Next, Paragraphs 19 through 22 of the Penney Reply describes a series of analyses of the single peak May model against other single peak models, concluding that the June model (with a May peak) is better than the others. These analyses are flawed in numerous ways.

29. First and most importantly, the entire exercise is based on a mischaracterization of my critique that implicitly assumes I am proposing a model with an April peak. I merely stated that a *naive* model such as the Penney Model could also be used to “prove” an April peak, meaning that such an analysis could also lead to spurious statistical significance. None of the Penney Reply analyses question this fact. I am not proposing that the data experienced a single change that caused the trend to abruptly reverse after April 2013 (a peak in that month and a decline thereafter). As I have stated numerous times, the data do not indicate a single change model is appropriate, whether that single change is in June 2013 or in some other month.

30. Second, in paragraph 19 of the Penney Reply, Dr. Penney attempts to complete a cross-validation analysis that uses three data sets for each of these article sets. However, two of the three models proposed in paragraph 19 of the Penney Reply, the “total page view” model and the “average total page view” model, are exactly the same statistically.³⁹ The total page view is simply the average page views multiplied by the number of articles. These two models are equivalent,

³⁷ Ibid, p. 6.

³⁸ Ibid, p. 3-4.

³⁹ Penney Reply, paragraph 19.

statistically, since regression models are invariant to changes in units.⁴⁰ For example, suppose we were trying to predict how far a person can jump according to their height in feet, and we ran a regression model that predicted someone who is 6 feet tall can jump 10 feet on average. If we use the same data but run the regression model based on inches, that new model would predict that someone who is 72 inches tall can jump 120 inches on average – in other words, the prediction is unchanged except for the expression in inches instead of feet.

31. The same is the case with running one model on the total and a second on the average, as is done in the Penney Reply (the results of which are summarized in the Penney Reply, paragraph 22). The model is unchanged but one is in terms of averages and one is in terms of totals. Therefore, the estimates for the model run on totals will be 23 times the estimates for the model run on the averages (for the Penney Reply model that has 23 articles). Thus, for example, in the Appendix to the Penney Reply showing the “23 Most Privacy Sensitive Article Set Cross Validation Analysis” (page 41), the coefficient for the variable *time* for the total model is shown to be 21,383.58. Two pages later (page 43), the same coefficient for the variable *time* in the average model is 929.72, which is exactly 21,383.58 divided by 23. The summary statistics like the t-statistic, which is 5.30, are also exactly the same.⁴¹ The Root Mean Square Error and Mean Absolute Errors highlighted for the total model are 89,506.35 and 63,503.27 (on page 41), which, when divided by the 23 articles considered, is equal to the highlighted totals of 3,891.54 and 2760.94 shown for the average model for the highlighted RMSE and Mean Absolute Error, respectively, shown in the attachments to the Penney Reply (on page 43).⁴²
32. Thus, while the Penney Reply asserts that there are 48 models (3 models by 4 datasets by 4 change points), there are really only 32 (2 models by 4 datasets by 4 change points). The four datasets also largely overlap, since the 46 article dataset includes all 44 articles in the 44 article dataset, which includes all 23 articles in the 23 article dataset, which includes all 21 articles in the 21 article dataset. In addition, the four months modeled are adjacent, meaning the regression models are very similar (this was part of my original point that the specification of the change point does not make much difference). In other words, though the Penney Reply asserts there are 48 separate models, there are only 32, and most of the 32 are highly related to one another and must produce similar results.
33. Third, the Penney Reply’s use of cross validation is misplaced and performed incorrectly. In part the Penney Reply employs a “cross validation analysis.”⁴³ This approach, which the Penney Reply uses to delete different time periods one at a time, is improper for time series models, in which the data points are related to one another.⁴⁴ In addition, the Penney Reply’s cross

⁴⁰ See, for example, Wooldridge, Jeffrey M., *Introductory Econometrics, A Modern Approach*, 5th Edition, 2012, South-Western Cengage Learning, p. 40-41.

⁴¹ The r-squared and the p-values are also exactly the same.

⁴² There is a slight difference due to rounding or less than 1 for each of the figures.

⁴³ Penney Reply, paragraph 19.

⁴⁴ This is because the data in the cross validation set, or the data “left out”, is not independent of the other data. See for example, Bergmeir, Christopher, and Benitez, Jose M., “On the use of cross-validation for time series predictor evaluation,” *Information Sciences*, 2012, 192-213. This paper discusses some of the fundamental problems with traditional cross-validation in time series, primarily in Sections 3.3 and 3.4. Also, see David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig and Carsten F. Dormann, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography* 40: 913-929 (913-925 in particular), 2017.

validation's purpose is to compare the June model to models with a different change point. None of the Penney Reply's cross validation analyses compare the simple single-change model to a model that accounts for other factors or otherwise corrects for omitted variables. Therefore, the Penney Reply's use of cross-validation to compare models and the attempt to show a May Peak model is better than an April peak model or other models are mere distractions that are not related to my criticism.

D. Data Beyond 2014 Show Article Views at About 2012 through 2014 Levels, Even When Earlier Data is Corrected for Mobile Views

34. In my February Declaration I pointed out that the extended view of page view data also does not indicate any long term decline. The Penney Reply, in paragraphs 23 and 34(g), responds to point out that my extended data includes mobile use while the original data presented in the Penney December Declaration did not. To address this "apples to oranges" comparison, I therefore adjusted the 2012 to 2014 data to account for mobile usage. As I describe below, this adjustment has no effect on my conclusions.
35. I also considered the effect that the non-inclusion of mobile usage and the lack of adjustment of that increasing usage had on the Penney Model. I find that such exclusion and lack of adjustment are additional flaws in the Penney Model. Specifically, the Penney December Declaration data excluded mobile page views from the terror and control data sets.⁴⁵ If these views were a constant percentage of total views, such an exclusion would likely not affect the Penney Model. However, as I explain below, there was a dramatic increase in mobile web access from January 2012, the first month of data included in the Penney December Declaration analysis, to August 2014, the last month included.
36. The data provided with the Penney December Declaration (but not used in the Penney December Declaration or the Penney Reply) indicates that in January 2012, mobile views accounted for about 12% of total page views.⁴⁶ By the end of the study period, that figure was 32%. In other words, the Penney December Declaration's exclusion of mobile views had an increasingly downward bias on total page views. This is yet another bias that affects the Penney Model, and, by not accounting for mobile visits, the Penney Model is biased toward finding an effect and toward finding a larger effect. This bias is a result of the fact that for later data the model excluded more views than for earlier data.⁴⁷
37. In terms of my graphs of extended data as compared to earlier data, the data prior to August 2014 would be higher with mobile data. My graphs included the data as originally provided with the Penney December Declaration, which did not include mobile data for the terror articles. Using

⁴⁵ It may be that mobile views were not available, in which case an adjustment, like the one I made, could have been made; or the Penney Model could have included a factor that accounts for such usage.

⁴⁶ This is based on the difference between the global English page views non-mobile and the total global English page views, and is consistent with this article <https://techcrunch.com/2016/11/01/mobile-internet-use-passes-desktop-for-the-first-time-study-finds/>.

⁴⁷ While it may seem that simply using desktop views only would not cause a bias to the results, this notion is not correct. To the extent that 1) mobile use was growing during the period, and 2) individuals were using mobile instead of (rather than in addition to) desktop views, the desktop views would be depressed in the latter part of the period and thus bias the results. This has occurred to such an extent that an increasing number of people rely exclusively on mobile access. See, for example, <https://techcrunch.com/2016/11/01/mobile-internet-use-passes-desktop-for-the-first-time-study-finds/>.

the Penney December Declaration's global article view dataset, which provides total views as well as total views excluding mobile, I adjusted the page views for the terror articles from January 2012 through August 2014 to account for mobile views.⁴⁸ The graphs below, showing the extended average and median page views with mobile page views factored in, are consistent with my earlier graphs of the extended data in that they indicate there was no downward trend after June 2013.⁴⁹

38. Average and median page views appear to decline some in mid-2017 but views in 2015 and 2016 appear to be at or above 2012 through 2014 levels. It is also notable in these longer data series that there are clear peaks around the times of major U.S. or European terror attacks, adding further evidence that any reasonable model would account for such attacks (and of course the Boston Marathon bombings occurred very close to the time of the alleged drop due to the Snowden disclosures).

⁴⁸ This rough adjustment is undoubtedly inaccurate but captures the magnitude and pattern of the mobile views.

⁴⁹ The adjustment results in an increase in article views for each month from January 2012 through August 2014, with the amount of increase depending on the share of total Wikipedia views that were mobile.

Figure 3: Average Page Views, Adjusting Data Before 2015 to Factor in Mobile Page Views

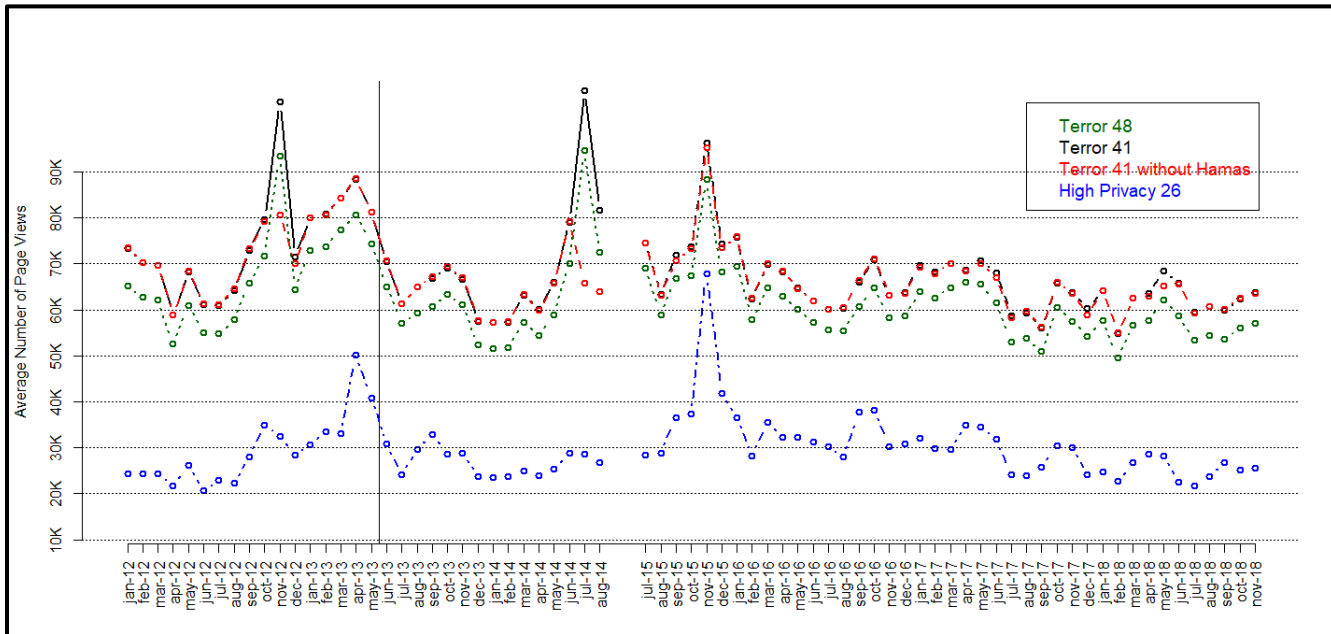
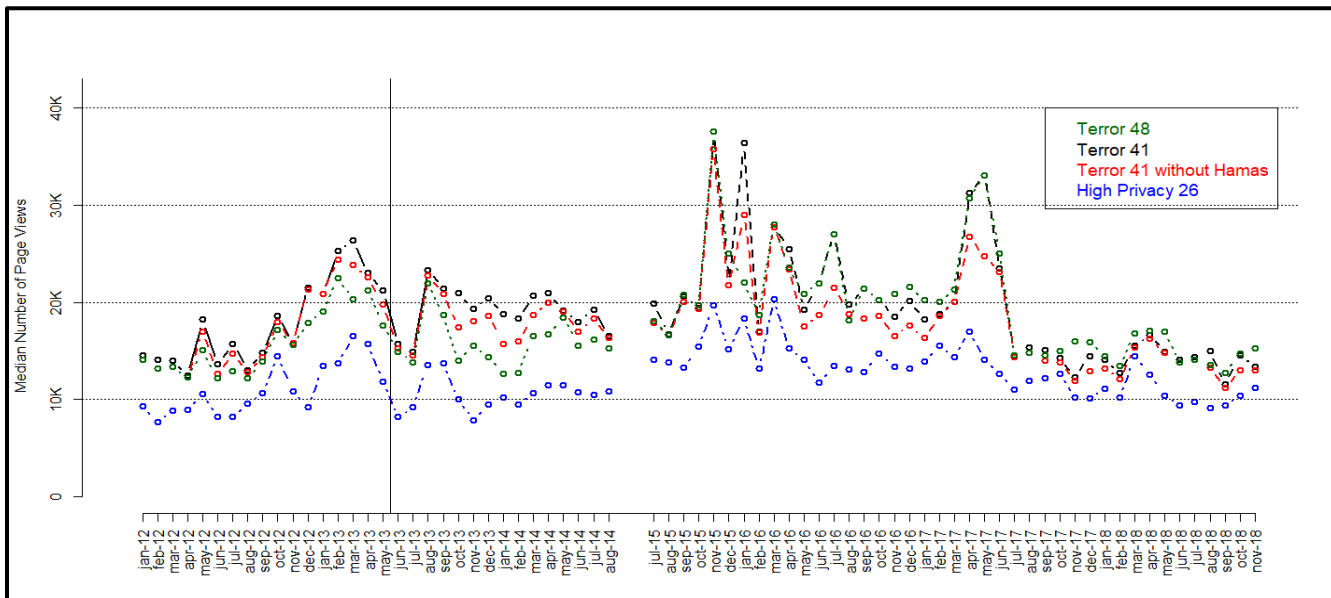


Figure 4: Median Page Views, Adjusting Data Before 2015 to Factor in Mobile Page Views



39. The Penney Reply cites some studies that purport to support the idea that the Penney December Declaration conclusions would continue beyond August 2014, but the Penney Reply neither considers (nor produced in this case) the data underlying those other studies. Even if those studies were to be based on a solid scientific and statistical grounds (and I cannot evaluate whether this is true without the underlying data), they only claim to offer conclusions applicable

to (at the latest) 2015.⁵⁰ Moreover, only one of the studies Dr. Penney cites in his reply appears to look at web data, rather than interview answers, and that study is from a working paper that was not published in a scientific journal and it expressly states it only includes data from 2013, and thus does not include any extended data. In any case, there is no way for me to evaluate the validity of those results, because I was not provided the data and it is not publicly available.

40. I do note that one article⁵¹ cited in the Penney Reply footnote 44 adjusts for additional variables and appears to find a smaller (and not statistically significant) effect in terms of drops in searches. This finding is consistent with omitted variable bias I outlined in my first Declaration with respect to the Penney December Declaration.⁵²

E. Omitted Variable Bias of the Penney Model Cannot be Solved by Deleting Valid Data.

41. In my February declaration, I pointed out a number of omitted variables that cause bias to the estimates made in the Penney December Declaration. These variables include ones associated with seasonality, individual differences in articles, and news events (the Boston Marathon bombings in particular).⁵³ The Penney Reply leaves these largely unaddressed but does assert it controls for seasonality because it includes more than one year of data before and after June 2013.⁵⁴ However, despite having sufficient data (barely), the Penney Model makes no correction for seasonality and includes no analysis that shows there is not such an effect. I showed such seasonal changes appear in this data and they are statistically significant.⁵⁵ In other words, though there was sufficient data, and that data shows statistically significant seasonal effects, the Penney December Declaration ignored seasonality. Wikimedia acknowledged these effects during the deposition of its designee, James Alexander: “global user base, especially in English Wikipedia, tends to have a bit of a dip during the summer, just because there are people out of school, and a lot of people use it in school or when they are studying.”⁵⁶ Curiously, the Penney Reply, paragraph 30a, states that there is “no basis to expect large seasonal effects with these page views.” This statement is speculation that flies in the face of the qualitative and statistical evidence.
42. The Penney declaration correctly states that “in a naturalistic study outside the experimental context, it is not possible to control for all confounding factors.”⁵⁷ However, the Penney December Declaration corrects for no confounding factors. As one recent author put it: “Obviously, one cannot include in a regression every variable that might conceivably be relevant. But when a factor has a reasonable chance of being important, to exclude it from the modeling is to risk substantial distortion.”⁵⁸ The Penney Reply re-asserts that the comparator datasets help

⁵⁰ Penney Reply, paragraph 34.

⁵¹ Section 3.2 of the article Marthews, Alex, and Tucker, Catherine, “Government Surveillance and Internet Search Behavior,” February 17, 2017, found at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2412564.

⁵² See p. 38 in Marthews and Tucker for lack of statistical significance. For my discussion of omitted variable bias, see Salzberg February Declaration, paragraphs 55-60.

⁵³ Salzberg February Declaration, paragraphs 55-60.

⁵⁴ Penney Reply, paragraph 30a.

⁵⁵ Salzberg February Declaration, paragraph 57.

⁵⁶ Deposition of Wikimedia designee, James Alexander, April 12, 2018, p. 145.

⁵⁷ Penney Reply, paragraph 30e.

⁵⁸ Barnett, Arnold I., Applied Statistics: Models and Intuition, Dynamic Ideas LLC, 2015, p. 582.

control for confounding factors, but this is not correct for article-specific factors and is not true when the comparator data is not comparable, as I showed in my February Declaration.⁵⁹

43. In some of the re-analyses in the Penney Reply, articles or time periods are deleted and the Penney Model is re-run.⁶⁰ It may be that the Penney Reply does this to address particular examples of the data not following the Penney Model. However, as I said above, these re-analyses do not support the results any more than the original analysis in the Penney December Declaration. Furthermore, by deleting data that tends to disprove the Penney Model and then re-running that data rigs the results toward adoption of the flawed Penney Model.
44. The Penney Reply seems to misinterpret my remarks concerning the staleness of a 2011 DHS list.⁶¹ I was not commenting on the objective nature of the selection, but rather that any list gets stale over time, and the list here used is no exception. For that reason, the static list has no mechanism to update the key articles and therefore a natural decline occurs. The same was not true for the comparator list of popular articles. Because the determination of which articles were popular was made after the time period studied in the Penney December Declaration, articles such as Deaths in 2014 -- which had virtually no page views in 2012 -- were part of the list.⁶² On the other hand, a group like ISIL/ISIS, which gained prominence in 2014, was not on the 2011 list, as I pointed out.⁶³

F. The Penney Model's Failure to Isolate the Effect of Public Awareness about the NSA Upstream Program

45. My sixth critique, discussed in my February declaration, is that "there are no data or statistical analysis offered that indicate such an effect [an abrupt decline in page views] was due to awareness of the specific NSA program at issue here rather than other related or unrelated events of June 2013."⁶⁴ The Penney Reply acknowledges that "in any study of naturalistic changes in human behavior, it will not be possible to isolate the source of all causes and effects on behavior" and that my critique is "a general observation about a [*sic*] naturalistic studies."⁶⁵ While this is correct, the Penney December Declaration analysis does not adjust for *any* of those causes, even the obvious ones like seasonality that affect summer page views.
46. Furthermore, the fact that the Penney Model may have been doomed from the start in terms of isolating the effect it intended to prove is not a reason for accepting the model; rather, it is a reason for rejecting it. Despite the passage of nearly six years since the Snowden disclosures, the Penney Reply does not cite a single study published in a peer-reviewed scientific journal that demonstrates the particular effect or even any chilling effect on Internet usage due to awareness of the actual operation of NSA programs.

⁵⁹ Salzberg February Declaration, paragraph 32-46.

⁶⁰ In Figure 3b and its explanation in the Penney Reply, the key month of April 2013 is deleted. In Figure 4 and accompanying analyses in the Penney Reply, eight of the original 31 high-privacy articles are deleted for reasons that are unclear to me and unstated in the Penney Reply.

⁶¹ Penney Reply, paragraphs 31 and 32, refer to this critique.

⁶² Salzberg Paragraph 64 and database showing 26 most popular articles, which accompanied the Penney December Declaration.

⁶³ Salzberg Declaration, paragraph 63.

⁶⁴ Salzberg Declaration, paragraph 66.

⁶⁵ Penney Reply paragraph 36a and 36c, for the first and second quoted material, respectively.

IV. Conclusion

47. In conclusion, my original critiques, detailed in my February Declaration are unchanged by the Penney Reply. In short, the analysis in the Penney December Declaration and the Penney Reply fail to show that public awareness of the Snowden revelations regarding the NSA Upstream program caused any drop in page views of Wikipedia articles.

I declare of penalty of perjury that the foregoing is true and correct to the best of my knowledge and belief. Executed in New York, New York on March 22, 2019.

A handwritten signature in black ink, appearing to read 'Alan J. Salzberg', is written over a horizontal line.

Alan J. Salzberg, Ph.D.
March 22, 2019

Appendix: Stata Program Log

The following log shows the results of the analysis I performed and described in this declaration. The program was run using Stata, Version 14.

```

name: <unnamed>
log:
D:\clients_2018\DOJ_Wiki_NSA\programsdata\penneyreply\regression_effects_20190318.log
log type: text
opened on: 19 Mar 2019, 11:01:40

. clear

.
. insheet using orig48long.csv
(23 vars, 3,504 obs)

. drop if artnames==" Hamas"
(73 observations deleted)

. save orig48long, replace
file orig48long.dta saved

. keep if monthindex<=32
(1,927 observations deleted)

. save orig48long32, replace
file orig48long32.dta saved

.
. *
. * Simulation that shows no difference in agg v. disagg if same model is run and
issue is just noise
. *
. use orig48long32, clear

. drop if artnames==" Hamas"
(0 observations deleted)

. drop if monthindex>32
(0 observations deleted)

. * run regression to get forecast error
. * no need to show output (but will show output of this for a different purpose
below)
. regress pageviews monthindex intervention postslope, noheader notable

. predict pviewmont
(option xb assumed; fitted values)

. predict sf, stdf

. * simulate data with same forecast error and run regression on disagg
. sort artnum monthindex

. isid artnum monthindex

. * set rndnum seed so can be replicated
. set seed 20190318

```

```
. gen errsim=rnormal(0,sf)

. replace pviewmont=pviewmont+errsim
(1,504 real changes made)
```

```
. regress pviewmont monthindex intervention postslope
```

Source	SS	df	MS	Number of obs	=	1,504
-----+				F(3, 1500)	=	3.35
Model	6.2583e+10	3	2.0861e+10	Prob > F	=	0.0185
Residual	9.3476e+12	1,500	6.2317e+09	R-squared	=	0.0067
-----+				Adj R-squared	=	0.0047
Total	9.4102e+12	1,503	6.2609e+09	Root MSE	=	78941

pviewmont	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+					
monthindex	568.2692	570.0658	1.00	0.319	-549.9417 1686.48
intervention	-11618.79	8230.636	-1.41	0.158	-27763.56 4525.991
postslope	-1155.076	893.594	-1.29	0.196	-2907.903 597.75
_cons	51521.53	5841.437	8.82	0.000	40063.28 62979.78

```
. * now aggregate, and see that regression standard errors and pvalues are about the
same
. * coefficients are exactly the same except for rounding because they do not depend on
simulation
. * the Root mean square error is about rmse of disagg model * sqrt(47), or about 7
times as high as mean
. collapse (mean) pviewmont , by( monthindex intervention postslope)
```

```
. regress pviewmont monthindex intervention postslope
```

Source	SS	df	MS	Number of obs	=	32
-----+				F(3, 28)	=	3.18
Model	1.3316e+09	3	443853226	Prob > F	=	0.0392
Residual	3.9062e+09	28	139508526	R-squared	=	0.2542
-----+				Adj R-squared	=	0.1743
Total	5.2378e+09	31	168961239	Root MSE	=	11811

pviewmont	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+					
monthindex	568.2691	584.7501	0.97	0.339	-629.5372 1766.075
intervention	-11618.78	8442.648	-1.38	0.180	-28912.77 5675.196
postslope	-1155.076	916.6119	-1.26	0.218	-3032.671 722.5181
_cons	51521.53	5991.905	8.60	0.000	39247.67 63795.39

```
.
. *
. * END Simulation
. *
.
.
. use orig48long32, clear

. drop if artnames=="Hamam"
(0 observations deleted)

. *
. * large changes in standard errors and stat. sign. with removal of a single
observation is another sign of a poor model
```

```

. *
. preserve

. keep if highprivind==1
(512 observations deleted)

. collapse (median) pageviews, by( monthindex intervention postslope highpriv)

. regress pageviews monthindex intervention postslope if highpriv==1

-----+-----
Source |          SS          df          MS      Number of obs   =          32
-----+-----
Model | 13595332.7            3  4531777.56      F(3, 28)         =          2.98
Residual | 42544868.8           28  1519459.6      Prob > F         =          0.0482
-----+-----
Total | 56140201.5           31  1810974.24      R-squared        =          0.2422
                                           Adj R-squared    =          0.1610
                                           Root MSE        =          1232.7

-----+-----
pageviews |          Coef.      Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
monthindex | 123.6005      61.02594      2.03  0.052     -1.405487    248.6065
intervention | -1336.267    881.0953     -1.52  0.141    -3141.109    468.5747
postslope | -189.8362    95.65985     -1.98  0.057    -385.7865    6.114118
_cons | 6285.478     625.3298     10.05  0.000     5004.548    7566.408

. restore

. preserve

. * possible error since recruitment and fundamentalism have exact same page views
nearly every month
. * thus show results without as well as with
. drop if artnames=="Recruitment" | artnames=="Fundamentalism"
(64 observations deleted)

. keep if highprivind==1
(480 observations deleted)

. collapse (median) pageviews, by( monthindex intervention postslope highpriv)

. regress pageviews monthindex intervention postslope if highpriv==1

-----+-----
Source |          SS          df          MS      Number of obs   =          32
-----+-----
Model | 9185572.85            3  3061857.62      F(3, 28)         =          4.55
Residual | 18850621.1           28  673236.47      Prob > F         =          0.0102
-----+-----
Total | 28036194           31  904393.355      R-squared        =          0.3276
                                           Adj R-squared    =          0.2556
                                           Root MSE        =          820.51

-----+-----
pageviews |          Coef.      Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
monthindex | 93.56127     40.62129      2.30  0.029     10.35233     176.7702
intervention | -1379.924    586.492     -2.35  0.026    -2581.298    -178.5493
postslope | -117.9791     63.675     -1.85  0.074    -248.4115    12.45319
_cons | 6070.125     416.2444     14.58  0.000     5217.487     6922.763

. restore

. preserve

```

```
. collapse (median) pageviews, by( monthindex intervention postslope)
```

```
. regress pageviews monthindex intervention postslope
```

Source	SS	df	MS	Number of obs	=	32
Model	84545042.9	3	28181681	F(3, 28)	=	8.18
Residual	96510363	28	3446798.68	Prob > F	=	0.0005
				R-squared	=	0.4670
				Adj R-squared	=	0.4098
Total	181055406	31	5840496.96	Root MSE	=	1856.6

pageviews	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
monthindex	374.8799	91.9132	4.08	0.000	186.6043	563.1556
intervention	-3299.076	1327.047	-2.49	0.019	-6017.408	-580.7433
postslope	-535.3763	144.0765	-3.72	0.001	-830.5036	-240.249
_cons	9601.022	941.83	10.19	0.000	7671.771	11530.27

```
. restore
```

```
. * now without possibly error data
. drop if artnames=="Fundamentalism" | artnames=="Recruitment"
(64 observations deleted)
```

```
. collapse (median) pageviews, by( monthindex intervention postslope)
```

```
. regress pageviews monthindex intervention postslope
```

Source	SS	df	MS	Number of obs	=	32
Model	72354244.3	3	24118081.4	F(3, 28)	=	17.36
Residual	38905201.2	28	1389471.47	Prob > F	=	0.0000
				R-squared	=	0.6503
				Adj R-squared	=	0.6129
Total	111259446	31	3589014.37	Root MSE	=	1178.8

pageviews	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
monthindex	315.9363	58.35724	5.41	0.000	196.3969	435.4757
intervention	-4298.331	842.5644	-5.10	0.000	-6024.246	-2572.416
postslope	-342.8184	91.47658	-3.75	0.001	-530.1997	-155.4371
_cons	8841.338	597.9838	14.79	0.000	7616.424	10066.25

```
. *
. * demonstrate that errors are correlated with articles, meaning disgregation or
some type of adjustment is needed
. * Also shows that stat significance does not exist for overall data
. use orig48long32, clear
```

```
. regress pageviews monthindex intervention postslope if artnames!="Fundamentalism" &
artnames!="Recruitment"
```

Source	SS	df	MS	Number of obs	=	1,440
Model	6.7546e+10	3	2.2515e+10	F(3, 1436)	=	3.37
Residual	9.5866e+12	1,436	6.6759e+09	Prob > F	=	0.0178
				R-squared	=	0.0070
				Adj R-squared	=	0.0049
Total	9.6541e+12	1,439	6.7089e+09	Root MSE	=	81706

pageviews	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
monthindex	752.6646	603.0018	1.25	0.212	-430.1942	1935.523
intervention	-14970.22	8706.167	-1.72	0.086	-32048.39	2107.947
postslope	-1179.932	945.2219	-1.25	0.212	-3034.096	674.2313
_cons	49658.62	6178.93	8.04	0.000	37537.93	61779.32

. regress pageviews monthindex intervention postslope

Source	SS	df	MS	Number of obs	=	1,504
Model	7.4228e+10	3	2.4743e+10	F(3, 1500)	=	3.86
Residual	9.6056e+12	1,500	6.4037e+09	Prob > F	=	0.0091
				R-squared	=	0.0077
				Adj R-squared	=	0.0057
Total	9.6798e+12	1,503	6.4403e+09	Root MSE	=	80023

pageviews	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
monthindex	881.2874	577.8797	1.53	0.127	-252.2505	2014.825
intervention	-14757.81	8343.452	-1.77	0.077	-31123.88	1608.265
postslope	-1436.449	905.8423	-1.59	0.113	-3213.301	340.4031
_cons	48705.37	5921.504	8.23	0.000	37090.07	60320.68

. predict residual1, residual

. * stat sign correlation between articles and residuals mean model is insufficient (see p-value and f-statistic)

. anova residual1 artnum

Number of obs = 1,504 R-squared = 0.9258
 Root MSE = 22124.8 Adj R-squared = 0.9234

Source	Partial SS	df	MS	F	Prob>F
Model	8.892e+12	46	1.933e+11	394.91	0.0000
artnum	8.892e+12	46	1.933e+11	394.91	0.0000
Residual	7.132e+11	1,457	4.895e+08		
Total	9.606e+12	1,503	6.391e+09		

. * note same coefficients in agg results

. collapse (mean) pageviews, by(monthindex intervention postslope)

. regress pageviews monthindex intervention postslope

Source	SS	df	MS	Number of obs	=	32
Model	1.5793e+09	3	526437311	F(3, 28)	=	24.85
Residual	593272771	28	21188313.2	Prob > F	=	0.0000
				R-squared	=	0.7269
				Adj R-squared	=	0.6977
Total	2.1726e+09	31	70083377.5	Root MSE	=	4603.1

pageviews	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
monthindex	881.2874	227.8862	3.87	0.001	414.4836	1348.091
intervention	-14757.81	3290.232	-4.49	0.000	-21497.54	-8018.073
postslope	-1436.449	357.218	-4.02	0.000	-2168.177	-704.721

```

-----
      _cons |   48705.37   2335.139   20.86   0.000   43922.06   53488.69
-----

```

```

.
. *
. * show art is also stat sign for 31 high privacy
. use orig48long32, clear

. keep if highpriv==1
(512 observations deleted)

. regress pageviews monthindex intervention postslope if artnames!="Fundamentalism" &
artnames!="Recruitment"

```

```

-----
      Source |          SS          df           MS      Number of obs   =          960
-----+-----
      Model | 1.3198e+10           3   4.3994e+09      F(3, 956)         =          4.20
      Residual | 1.0017e+12          956   1.0478e+09      Prob > F           =         0.0058
-----+-----
      Total | 1.0149e+12          959   1.0582e+09      R-squared           =         0.0130
      Root MSE |                               =         0.0099
      Adj R-squared =         0.0099
      Root MSE   =         32369

```

```

-----
      pageviews |          Coef.      Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      monthindex |    823.6276     292.5762      2.82  0.005     249.4618     1397.793
      intervention |   -8112.912    4224.229     -1.92  0.055    -16402.74     176.9203
      postslope |   -1145.897    458.6213     -2.50  0.013    -2045.918    -245.8766
      _cons |    14796.3     2998.014      4.94  0.000     8912.854     20679.75
-----

```

```

. regress pageviews monthindex intervention postslope

```

```

-----
      Source |          SS          df           MS      Number of obs   =          992
-----+-----
      Model | 1.6582e+10           3   5.5273e+09      F(3, 988)         =          5.18
      Residual | 1.0532e+12          988   1.0660e+09      Prob > F           =         0.0015
-----+-----
      Total | 1.0698e+12          991   1.0795e+09      R-squared           =         0.0155
      Adj R-squared =         0.0125
      Root MSE |                               =         0.0125
      Root MSE   =         32650

```

```

-----
      pageviews |          Coef.      Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      monthindex |    918.8429     290.3169      3.16  0.002     349.1343     1488.552
      intervention |   -8179.243    4191.609     -1.95  0.051    -16404.72     46.23536
      postslope |   -1340.458    455.0798     -2.95  0.003    -2233.492    -447.4243
      _cons |    15198.27    2974.863      5.11  0.000     9360.491     21036.04
-----

```

```

. predict residual1, residual

```

```

. * stat sign correlation between articles and residuals mean model is insufficient
(see p-value and f-statistic)
. anova residual1 artnum

```

```

      Number of obs =          992      R-squared      = 0.8558
      Root MSE      =    12570.6      Adj R-squared = 0.8513

```

```

-----
      Source | Partial SS          df           MS          F          Prob>F
-----+-----
      Model | 9.014e+11           30   3.005e+10     190.14   0.0000
      |
      artnum | 9.014e+11           30   3.005e+10     190.14   0.0000
      |

```


Residual		1.519e+11	961	1.580e+08

Total		1.053e+12	991	1.063e+09

```
. * note same coefficients in agg results
. collapse (mean) pageviews, by( monthindex intervention postslope)

. regress pageviews monthindex intervention postslope
```

Source		SS	df	MS	Number of obs	=	32

Model		534899840	3	178299947	F(3, 28)	=	20.87
Residual		239215204	28	8543400.16	Prob > F	=	0.0000

Total		774115045	31	24971453.1	R-squared	=	0.6910
					Adj R-squared	=	0.6579
					Root MSE	=	2922.9

pageviews		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

monthindex		918.8429	144.7056	6.35	0.000	622.4269	1215.259
intervention		-8179.243	2089.266	-3.91	0.001	-12458.91	-3899.577
postslope		-1340.458	226.83	-5.91	0.000	-1805.099	-875.8181
_cons		15198.27	1482.791	10.25	0.000	12160.91	18235.63

```
.
. log close
  name: <unnamed>
  log:
D:\clients_2018\DOJ_Wiki_NSA\programsdata\penneyreply\regression_effects_20190318.log
  log type: text
  closed on: 19 Mar 2019, 11:01:40
```