Sign-Congruence, External Validity, and Replication
For Online Publication

Tara Slough* Scott A. Tyson†

S1 Application II: Sports Outcomes and Pro-Incumbent Voting

Our framework is also useful for understanding replication efforts that employ observational data, where determining what constitutes a replication is less clear. To illustrate how to use our framework in such contexts, we consider an ongoing debate about whether sporting game outcomes affect pro-incumbent voting. We use this example because it comes from a lengthy published back-and-forth about how replication should be conducted in observational research, and thus provides a unique opportunity where issues about how to analyze and interpret replications with observational data are discussed in print.

In an important contribution, Healy, Malhotra, and Mo (2010) find that college football victories, which occur in the two weeks before general elections for president, governor, and senator, increase the incumbent party's vote share in the county where the university is located. The sample used in their study consists of presidential elections from 1960-2004, and gubernatorial and senate elections from 1967-2006. In terms of the mechanism, the authors posit that shocks to voter well-being (football victories) increase voter satisfaction with the status quo. Because the incumbent party represents the status-quo, this increased satisfaction translates into higher incumbent vote share. Healy, Malhotra, and Mo (2015, p. 12804) further elaborate this mechanism, writing:

> "Voters who are in a positive state of mind on Election Day are likely to use their mood as a signal for the incumbent party's success...and access positive memories about the incumbent party...and/or interpret past actions taken by the incumbent party more favorably...Additionally, positive emotions may cause voters to be more satisfied with the status quo...Those voters may then be more likely to choose the incumbent party in the election."

Since college football victories are thought to be outside the purview of presidents, governors, and senators, this finding—if it arises elsewhere—raises important questions about voter rationality and, as a result, the limits of democratic accountability.[1]

In response to the original Healy, Malhotra, and Mo (2010) paper, and sparking the debate about replication that interests us, Fowler and Montagnes (2015) argue that the finding that college football victories

*Assistant Professor, New York University. taraslough@nyu.edu
†Associate Professor, Emory University. s.tyson@emory.edu

[1]We do not contribute to the discussion of voter rationality, or what constitutes an "irrelevant event," for a discussion see Ashworth, Bueno de Mesquita, and Friedenberg (2017, 2018).

increase pro-incumbent voting is likely a false positive. Their argument is built upon a number of analyses. First, they extend the panel from the original Healy, Malhotra, and Mo (2010) study to presidential elections from 1960-2012 (i.e., adding 2004-2012), and gubernatorial and senate elections from 1960-2006 (i.e., adding 1960-1967). Using the extended sample, Fowler and Montagnes (2015) test a number of ancillary hypotheses that are consistent with Healy, Malhotra, and Mo (2010)'s proposed mechanism, and also include an alternative set of specifications with county-year fixed effects. In addition, Fowler and Montagnes (2015) conduct what is best described as a conceptual replication using NFL games, arguing that the mechanism proposed by Healy, Malhotra, and Mo (2010) should operate on such victories as well, especially since NFL games enjoy higher viewership and a more loyal following by fans. The additional specifications, additional data, and conceptual replication analyzed by Fowler and Montagnes (2015) ultimately do not recover evidence that is consistent with the findings originally reported in Healy, Malhotra, and Mo (2010). This lack of evidence led Fowler and Montagnes (2015) to conclude that there is no systematic evidence to support the argument that sporting outcome shocks, which may influence voter well-being, benefit incumbent politicians electorally.

In a direct response to Fowler and Montagnes (2015)'s critique, Healy, Malhotra, and Mo (2015) argue that Fowler and Montagnes (2015) do not conduct a true replication. Specifically, they claim that Fowler and Montagnes (2015) do not consider the totality of the evidence presented because they do not consider the survey evidence on NCAA basketball games that was discussed in Healy, Malhotra, and Mo (2015). Moreover, Healy, Malhotra, and Mo (2015) claim that Fowler and Montagnes (2015) neglect their preferred specification, which accounts for a team's *ex-ante* probability of victory, thereby isolating the effect of unexpected victories.[2]

Following up with a different set of replications, Graham et al. (2021) conduct a pre-specified replication of several studies of voter competence/rationality, including Healy, Malhotra, and Mo (2010). In addition to correcting several data errors in Healy, Malhotra, and Mo (2010) (see the supplemental information of Graham et al. (2021)), they extend the time series slightly. Their preferred specification pools the (corrected) in-sample data with the new (previously) out-of-sample data and show that estimates are attenuated, but in the same direction as the original finding. In a response, Fowler and Montagnes (2022*a*) argue that Graham et al. (2021) overstate the strength of evidence consistent with Healy, Malhotra, and Mo (2010)'s claims. In particular, they distinguish between in-sample and out-of-sample data, and conduct a simulation to show that the evidence on the pooled sample cannot reject (statistically) the possibility that the Healy, Malhotra, and Mo (2010) was a false positive.[3] Table S1 summarizes the published (or forthcoming) papers associated

[2]The authors' preferred operationalization of treatment measures a surprise football victory as:

$$W_{it} = \text{Win}_{it} - \Phi\left(\frac{-x}{13.89}\right),$$

where Win_{it} is a binary indicator that takes a value of 1 when the county's team wins game t; x is the game's points spread; and Φ is the standard normal cdf. They define this at different points (two weeks before the election, one week before the election, and both games). Note that $W_{it} \in (-1, 1)$, where -1 is a completely unexpected loss and 1 is a completely unexpected victory.

[3]Continuing the back-and-forth, Graham et al. (2022) criticize Fowler and Montagnes (2022*a*)'s treatment of multiple specifications, which weigh the results from different specifications equally. They instead argue for prioritization of average effects over

Citation	Summary
Healy, Malhotra, and Mo (2010)	Finds that college football victories in the two weeks before general elections for president, governor, and senator increase the incumbent party's vote share. The mechanism is shocks to voter well-being (football victories) increase voter-satisfaction with the status-quo, translating to increased incumbent vote share.
Fowler and Montagnes (2015)	Argues that HMM2010 is likely a false positve. They re-analyze the HMM2010 data using alternative specifications, conduct the HMM2010 analysis on a longer panel, as well as seeing if the same result holds also for NFL game outcomes. They do not find evidence consistent with the posited mechanism (shocks to voter well-being).
Healy, Malhotra, and Mo (2015)	Argues that FM2015 do not consider the totality of the evidence presented because they do not consider the survey evidence on NCAA basketball games or the preferred specfication that adjusts for the probability of victory.
Graham et al. (2021)	Conduct a pre-specified replication of voter competence/rationality including HMM2010, extending the original time series.† Their preferred specification shows that estimates are attenuated, but in the same direction as the original finding.
Fowler and Montagnes (2022*a*)	Argue that Graham et al. (2021) overstate the strength of evidence consistent with Healy, Malhotra, and Mo (2010), noting that they cannot reject (statistically) the possibility that the Healy, Malhotra, and Mo (2010) was a false positive.
Graham et al. (2022)	Contest equal treatment of multiple specifications and advocate for replication on an expanded sample that consists of both in-sample and out-of-sample observations.
Fowler and Montagnes (2022*b*)	Justifies the focus on multiple pre-specfied tests and argues for the merits of out-of-sample replication.

Table S1: Summary of replications and responses to Healy, Malhotra, and Mo (2010).
†: FM2022a note that GHMM2021 rely on a subset of the original data starting in 1985 rather than using the full (original) sample.

with this debate.

Fowler and Montagnes (2015, 2022*a*,*b*) all suggest that the original results in Healy, Malhotra, and Mo (2010), that sports outcomes affect voter assessment of incumbents, are likely false positives (Type-I errors). The responses of Healy, Malhotra, and Mo (2015), and Graham et al. (2021, 2022) claim that the results of Healy, Malhotra, and Mo (2010), updated in Graham et al. (2021) reflect a genuine effect. The substance of the debate, as it pertains to replication, centers on statistical questions, mostly about what constitutes the appropriate sample or the right regression specification. While we do not weigh in on the substantive debate, it is worth pointing out that this debate treats replication conceptually. Our framework clarifies some core disagreements between the two teams of scholars when thinking about replication. We will focus on two features of the debate that speak to issues of replication: (i) the presence (or lack thereof) of a similar effect with respect to NFL games; and (ii) the disagreement regarding the treatment of in- and out-of-sample data.

NFL versus NCAA victories: Does the effect of NFL game victories (expected or otherwise) on incumbent vote share constitute a replication of the Healy, Malhotra, and Mo (2010) finding that college football victories improve incumbent vote share? Recalling that an empirical target is denoted by τ, the key theoretical claim by Fowler and Montagnes (2015) is that:

$$\tau_{NFL} > \tau_{NCAA},$$

heterogeneous treatment effects.

i.e., NFL victories should have a larger (positive) effect on incumbent vote share than NCAA victories on incumbent vote share. They argue "we would expect NFL games to have a greater effect than college football games, because the NFL is significantly more popular—television ratings are $\sim$10 times greater—and NFL teams receive strong regional support just like college teams" (p. 13802-3). This argument supposes that NFL victories are a *stronger treatment*, which suggests that the hypothesized claim about NFL games arises as an *artifactual* discrepancy. But there could be target discrepancies as well. For example, may be the case that the voter mood mechanism produces different effects on different cross sections of counties. NFL teams tend to be located in larger metropolitan areas, on average, than NCAA teams. It could be the case that the effect of the mechanism depends on metro-area population (due to, for example, the availability of non-football related activities). While the justification for this claim invokes an expectation of artifactual discrepancies, the claim—and hence the test—is actually a statement about the sum of target and artifactual discrepancies.

Fowler and Montagnes (2022*b*) interpret the NFL replication in light of their expecation that artifactual discrepancies will be positive and target discrepancies are absent. Under this assumption, evidence that NFL games do not influence incumbent vote share suggests that NCAA games do not influence incumbent vote share. Although Healy, Malhotra, and Mo (2015), Graham et al. (2021), or Graham et al. (2022) do not explicitly address the lack of evidence of a similar effect from NFL games, there are two ways to interpret Fowler and Montagnes (2015)'s replication finding. First, it is possible that the mechanism endorsed by Healy, Malhotra, and Mo (2010, 2015) lacks external validity such that it is only true of the instance of college football games in their sample (offsetting the hypothesized positive artifactual discrepancy). Consequently, their result can be seen as a historical accident, and the normative concerns that motivate the substantive debate about voter rationality are less worrisome than they may otherwise be.

Second, Fowler and Montagnes (2022*b*)'s assumption that artifactual discrepancies are (weakly) positive for the NFL victories treatment could be wrong and negative artifactual discrepancies could instead be present. In this case, comparison of the sign of results would mislead inferences about the external validity of Healy, Malhotra, and Mo (2010)'s mechanism. In this interpretation, NFL games are not "similar enough" to provide an alternative measure of the mechanism at play in the NCAA result. As a result, one should not expect the same kind of effect for NFL games.

Finally, if we should not expect NFL games to yield a similar effect to NCAA games, then Healy, Malhotra, and Mo (2015)'s description of the mechanism (see above) is insufficient, since this description does not connect the mechanism to *college* football games, but to voter mood. It suggests that many things that positively effect a voter's mood should improve incumbent vote share, which seemingly should apply to NFL games. In sum, the inability to replicate the result with NFL games suggests either that the mechanism is ephemeral and does not generalize (Glennan, 2010), or that the mechanism studied is considerably narrower than Healy, Malhotra, and Mo (2015) and Graham et al. (2021) suggest.

In- versus out-of-sample: Graham et al. (2022) and Fowler and Montagnes (2022*b*) disagree on whether the original sample should be pooled with out-of-sample replication data. The primary argument of

Fowler and Montagnes (2015) is that the result suggesting that college football victories improve incumbent vote share is a false positive. To show this, they conduct an analysis similar to Healy, Malhotra, and Mo (2015) but on data that is outside the original sample. We note that they test whether treatment effects are different from zero but do not conduct a formal test comparing the original and replication estimates. Graham et al. (2021), instead, take the new data and *combine it with the original sample where the purported false positive is present*, to see if the result still maintains. They find that the original result is attenuated. Specifically, Graham et al. (2021) find a reduced influence of college football victories on incumbent vote share when combining additional data with the original sample. Graham et al. (2022) argue for pooling the original sample with new data, while Fowler and Montagnes (2022*b*) argue for the merits of out-of-sample replication and comparison of findings.

Why does the in- and out-of-sample definition matter when considering a replication? A replication is about a comparison between empirical targets (through estimates) that reflect different settings. Fowler and Montagnes (2015)'s analysis splits up the available data into two distinct samples that reflect different "settings" essentially making their exercise a replication assessing whether the empirical targets are similar, i.e., whether the voter mood mechanism has sign-congruent external validity. Graham et al. (2022)'s argument to pool all the data reflects a statistical concern, specifically, more data is better. However, their exercise much more closely resembles a meta-analysis, which *assumes* target-equivalence across the two settings. But there are reasons that one may be skeptical of this claim. For instance, college football viewership has held steady over the past 20 years,[4] whereas the population—and hence the pool of eligible voters—has grown. This renders college football victories a *weaker* treatment for the electoral application. We might, then, expect the effect of college football on incumbent vote share to attenuate toward zero since a smaller subset of the voting population is watching college football.

Before concluding, it is worth mentioning that statistical discrepancies are necessarily present. We know that the estimates using any of the above settings or measurement strategies will be measured with error. In the best case, Fowler and Montagnes (2015) test a null hypothesis of zero in different subsets of the data (with different specifications). Figure 4 in the main text of our paper shows how independent hypothesis tests of a null hypothesis of zero can lead to misleading inferences using heuristic versions of the sign-comparison test. Our estimate- or sign-comparison test both provide formal tests that can be applied in experimental and observational replications. In sum, our framework and approach to comparison of estimates in replication studies can be productively applied to observational studies.

[4]See, for example, https://www.sportsmediawatch.com/2021/01/national-championship-ratings-record-low-audience-alabama-ohio-state/.

Supplementary Appendix: References

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2017. "Accountability and information in elections." *American Economic Journal: Microeconomics* 9 (2): 95–138.

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2018. "Learning about voter rationality." *American Journal of Political Science* 62 (1): 37–54.

Fowler, Anthony, and B. Pablo Montagnes. 2015. "College football, elections, and false-positive results inobservational research." *Proceedings of the National Academy of Sciences* 112 (45): 13800–13804.

Fowler, Anthony, and B. Pablo Montagnes. 2022*a*. "Distinguishing between False Positives and Genuine Results: The Case of Irrelevant Events and Elections." *Journal of Politics* Forthcoming.

Fowler, Anthony, and B. Pablo Montagnes. 2022*b*. "On the Importance of Independent Evidence: A Reply to Graham et al." Working paper, available at `https://drive.google.com/file/d/16bV6Cyhau6spf6ahz4P1eO1k-O7lR2lt/view`.

Glennan, Stuart. 2010. "Ephemeral mechanisms and historical explanation." *Erkenntnis* 72 (2): 251–266.

Graham, Matthew H., Gregory A. Huber, Neil Malhotra, and Cecilia Hyunjung Mo. 2021. "Irrelevant Events and Voting Behavior:Replications Using Principles from Open Science." *Journal of Politics* Forthcoming.

Graham, Matthew H., Gregory A. Huber, Neil Malhotra, and Cecilia Hyunjung Mo. 2022. "How Should We Think About Replicating Observational Studies? A Reply to Fowler and Montagnes." *Journal of Politics* Forthcoming.

Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2010. "Irrelevant events affect voters'evaluations ofgovernment performance." *Proceedings of the National Academy of Sciences* 107 (29): 12804–12809.

Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2015. "Determining false-positives requires consideringthe totality of evidence." *Proceedings of the National Academy of Sciences* 112 (48): E6591.