SHORT REPORT

Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects

Samara Klar*, Thomas Leeper[†] and Joshua Robison[‡]

*School of Government and Public Policy, University of Arizona, Tucson, AZ 85721, USA, e-mail: Klar@email.arizona.edu, Twitter: @SamaraKlar, [†]Department of Methodology, London School of Economics and Political Science, London WC2A 2AE, UK, e-mail: thosjleeper@gmail.com, Twitter: @ThosJLeeper and [‡]Institute of Political Science, Leiden University, Leiden, Netherlands, e-mail: j.a.robison@fsw.leidenuniv.nl, Twitter: @JRob617

Abstract
Scholars from across the social sciences argue that identities – such as race, ethnicity, and gender – are highly influential over individuals' attitudes, actions, and evaluations. Experiments are becoming particularly integral for allowing identity scholars to explain how these social attachments shape our political behavior. In this letter, we draw attention to how identity scholars should approach the common practice of assessing moderators, measuring control variables, and detecting effect heterogeneity using covariates. Special care must be taken when deciding where to place measures of demographic covariates in identity-related experiments, as these cases pose unique challenges from how scholars traditionally approach experimental design. We argue in this letter that identity scholars, particularly those whose subjects identify as women or minorities, are often right to measure covariates of interest posttreatment.

Keywords: Identities; experiments; priming; experimental design; posttreatment bias

Experiments can help scholars to explain how individuals' identities shape their political behavior. We aim to draw attention to risks identity scholars face when placing identity-related covariates in an experimental design. Helpful work exists on this practice (e.g., Acharya, Blackwell and Sen 2016; Blackwell 2013; Montgomery, Nyhan and Torres 2018), but does not speak to the unique setting of identity politics, wherein key covariates can exert consequential priming effects.

Best practices for analyzing experiments warn against conditioning results on posttreatment covariates. As Acharya, Blackwell and Sen (2016) explain, "conditioning on a posttreatment variable changes the quantity of interest from an overall average

treatment effect to a direct effect of the treatment net the posttreatment variable" (514). Montgomery, Nyhan and Torres (2018) recommend measuring covariates before the treatment using a panel or "before the experimental manipulation during a single survey" (773). They state that "even variables that seem likely to remain fixed when measured after treatment, such as measures of racial or partisan identification, can be affected by treatments".[1]

Measuring covariates posttreatment can indeed create analytical problems. However, experimentalists should not always measure identities pretreatment either. Instead, researchers must base this decision on case-specific theory regarding the relationship between the treatment and measure of identity, and an explicit trade-off of the risks of posttreatment bias and priming effects.

Priming occurs when a consideration becomes accessible and receives extra weight when forming subsequent evaluations (e.g., Mendelberg 2008). Countless studies document the ease of priming and its consequences. Transue (2007) finds asking Americans about their national identity substantially influences their support for a tax increase (see also Sniderman, Hagendoorn and Prior 2004). Morris, Carranza and Fox (2008) ask respondents for their political identification either at the beginning or at the end of a survey. The former design leads Republicans to choose higher-risk investment preferences and Democrats to choose lower-risk preferences. Klar (2013) asked respondents to consider their partisanship before evaluating policies, finding that even the weakest partisan primes significantly change policy evaluations.

Identity primes can even contaminate attitudes that seem unrelated to identity-based interests. Williams et al. (2008) randomly asked respondents a question about their race either before or after completing a survey on public health attitudes. The prime influenced their responses about health and hygiene.[2] Benjamin, Choi and Strickland (2010) randomly assigned respondents to complete a questionnaire about either their linguistic/immigration history or their cable subscriptions. Asian subjects who received the linguistic/immigration questions displayed significantly different reward-seeking behaviors, as did African-American subjects in a replication.

Identification comprises both group membership (e.g., I am a woman) and the strength of that identification (e.g., My gender is important to me). Reports of group membership are highly stable and we know of no experiments that change a respondent's reported self-categorization regarding their gender, racial, or religious self-classification (i.e., from one category to another). While certain contexts can persuade Independents to report a partisan leaning (e.g., Craig and Richeson 2014; Klar and Krupnikov 2016) and the intensity with which partisans identify

---

[1]Montgomery, Nyhan and Torres (2018) cite Antman and Duncan (2015) as evidence for the instability of race. However, this study does not use experimental data. They cite Weiner (2015), meanwhile, as evidence for instability in partisanship. Here, respondents that answered questions about New Jersey's budgetary and regulatory response to superstorm Sandy were less likely to report being an Independent or possessing no partisan preference than those randomly assigned to answer questions about their personal experiences. The study does not show instability among partisans though, a point we return to later.

[2]Race scholars are particularly attuned to these concerns and often explicitly measure race/ethnic identity posttreatment to avoid contaminating their studies via priming (see for e.g. Jackson 2011; Mendelberg 2008; Perez et al. 2019).

Table 1
A Framework for Measurement Order Decisions

Qualities of covariate and treatment, as determined by the researcher	Pretreatment measure	Posttreatment measure
i. *Covariate is fixed/stable*	Risk of priming effects	OK
ii. *Covariate is susceptible to treatment* iii. *Covariate cannot contaminate study via priming effects*	OK	Risk of posttreatment bias
i. *Covariate is not susceptible to treatment* ii. *Covariate can contaminate study via priming effects*	Risk of priming effects	OK
i. *Covariate is susceptible to treatment* ii. *Covariate cannot contaminate study via priming effects*	Risk of priming effects	Risk of posttreatment bias

with their group is malleable (e.g., Abrams et al. 1990; Lupu 2013), we know of no treatments that lead partisans to change their party preference (e.g., from Democrat to Republican or vice versa). This is consistent with the tremendous stability of partisan classifications (Green, Palmquist and Schickler 2002). It is thus crucial to consider the specific type of identity measure when considering where to place identity-related covariates. If the treatment influences the measure, then posttreatment indicators bias estimates of the intended causal effect. Pretreatment measures, on the other hand, may change the definition of the causal parameter being estimated from the effect of the treatment when identity is non-salient to the effect when it is salient, which may not be the effect the experiment is interested in detecting.

One solution is to simply place covariates in a distinct wave using panel data. However, this is infeasible in many settings [e.g., exit polls (Klar 2013), rallies (McClendon 2014), or public events (Harrison and Michelson 2016)]. Panels are also prohibitively expensive: a two-wave survey of 1000 Americans is approximately three times more costly than a one-wave sample due to attrition, re-contact, and re-incentivizing respondents. Another proposed solution is to use a list of "buffer" items intended to dilute priming effects. However, primes can persist beyond such buffers (Druckman and Chong 2010). Finally, it is possible to simply check for posttreatment bias by regressing posttreatment measures on indicators for each treatment (e.g., Mummolo 2016). Yet, this modified form of a balance test also poses its own risks with respect to Type I error (see Mutz, Pemantle and Pham 2019).

We recommend that identity scholars explicitly base their design choice on case-specific theory concerning the relationship between the treatment and the measurement of the identity. Social identity approaches, for instance, can be helpful in considering cases when an identity may be affected by a treatment. Placing identity questions on the posttest is problematic when the treatment threatens the "value" or "distinctiveness" of the identity, as this can impact the self-categorization, and behavior, of high and low identifiers in offsetting ways (e.g., Branscombe et al. 1999). However, if the treatment does not target the value of an identity, and that identity has no demonstrated history of instability, then it is defensible to place it on the posttest. Table 1 provides a schematic for thinking through this process. It is up

to the researcher to theorize, based on existing literature and, if possible, pretests, whether (i) the covariate is susceptible to being influenced by the treatment and (ii) the covariate might contaminate the study via priming. Depending on each consideration, identity scholars should carefully decide whether covariates should appear pre- or posttreatment.

Demographic and social identities can be consequential in any setting. Identity scholars must consider the *causal order* of variables and the risks of each form of bias, rather than following conventional wisdom that might not apply to them.

References

Abrams, Dominic, Margaret Wetherell, Sandra Cochrane, Michael A. Hogg, and John C. Turner. 1990. Know What to Think by Knowing Who You Are: Self-Categorization and the Nature of Norm Formation, Conformity and Group Polarization. *British Journal of Social Psychology 29*(2): 97–119.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3): 512–29.

Antman, Francisca and Brian Duncan. 2015. Incentives to identify: Racial identity in the age of affirmative action. *Review of Economics and Statistics 97*(3): 710–3.

Benjamin, Daniel J., James C. Choi and A. Joshua Strickland. 2010. Social identity and preferences. *American Economic Review 100*(2): 1913–28.

Blackwell, Matthew. 2013. A framework for dynamic causal inference in political science. *American Journal of Political Science 57*(2): 504–19.

Branscombe, Nyla R, Naomi Ellemers, Russell Spears, and Bertjan Doosje. 1999. The Context and Content of Social Identity Threat. In *Social Identity*, eds. Ellemers N., Spears R. and Doosje B. Oxford: Blackwell Publishers, 35–58.

Craig, Maureen A., and Jennifer A. Richeson. 2014. On the precipice of a "majority-minority" America: Perceived status threat from the racial demographic shift affects white Americans' political ideology. *Psychological Science 25*(6): 1189–97.

Druckman, James N., and Dennis Chong. 2010. Dynamic public opinion: Communication effects over time. *American Political Science Review 104*(4): 663–80.

Green, Donald P., Bradley Palmquist and Eric Schickler. 2002. *Partisan Hearts and Minds*. Yale, New Haven: Yale University Press.

Harrison, Brian F., and Melissa R. Michelson. 2016. More than a game: Football fans and marriage equality. *PS: Political Science & Politics 49*(4): 782–7.

Jackson, Melinda. 2011. Priming the sleeping giant: The dynamics of latino political identity and vote choice. *Political Psychology 32*(4): 691–716.

Klar, Samara. 2013. The influence of competing identity primes on political preferences. *The Journal of Politics 75*(4): 1108–24.

Klar, Samara, and Yanna Krupnikov. 2016. *Independent Politics: How American Disdain for Parties Leads to Political Inaction*. New York, NY: Cambridge University Press.

Lupu, Noam. 2013. Party brands and partisanship: Theory with evidence from a survey experiment in Argentina. *American Journal of Political Science 57*(1): 49–64.

McClendon, Gwyneth. 2014. Social esteem and participation in contentious politics: A field experiments at an LGBT pride rally. *American Journal of Political Science 58*(2): 279–90.

Mendelberg, Tali. 2008. Racial priming revived. *Perspectives on Politics 6*(1): 109–23.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. How condition on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science 62*(3): 760–75.

Morris, Michael W., Erica Carranza, and Craig R. Fox. 2008. Mistaken Identity: Activating Conservative Political Identities Induces Conservative Financial Decisions. *Psychological Science 19*(11): 1154–60.

Mummolo, Jonathan. 2016. News from the other side: How topic relevance limits the prevalence of partisan selective exposure. *The Journal of Politics 78*(3): 763–73.

Mutz, Diana, Robin Pemantle, and Philip Pham. 2019. The Perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician 73*(1): 32–42.

Pérez, Efrén O., Andrew M. Engelhardt, and Maggie Deichert. 2019. E Pluribus Unum? How ethnic and national identity motivate reactions to a political ideal. *The Journal of Politics. 81*(4). doi: 10.1086/704596.

Sniderman, Paul M., Louk Hagendoorn, and Markus Prior. 2004. Predisposing factors and situational triggers: Exclusionary reactions to immigrant minorities. *American Political Science Review 98*(1): 35–49.

Transue, John E. 2007. Identity salience, identity acceptance, and racial policy attitudes: American national identity as a uniting force. *American Journal of Political Science 51*(1): 78–91.

Weiner, Marc D. 2015. A Natural experiment: Inadvertent priming of party identification in a split-sample survey. *Survey Practice 8*(6): 1–7.

Williams, Monnica T., Eric Turkheimer, Emily Magee, and Thomas Guterbock. 2008. The effects of race and racial priming on self-report of contamination anxiety. *Personality and Individual Differences 44*(3): 746–57.