Observational Equivalence in Explaining Attitude Change: Have White Racial Attitudes Genuinely Changed?*

Running header: Observational Equivalence in Explaining Attitude Change

Keywords: racial attitudes, measurement equivalence, attitude change, polarization

Andrew M. Engelhardt
amengelhard@uncg.edu
The Department of Political Science
University of North Carolina, Greensboro
324 Curry Building
PO Box 26170
Greensboro, NC 27402

* I thank Efrén Pérez, the Brown Political Behavior Workshop, the editors and 3 reviewers for helpful comments and feedback.

Observational Equivalence in Explaining Attitude Change: Have White Racial Attitudes Genuinely Changed?

Abstract

Understanding when and why White racial attitudes change is important for understanding their politics. Critically, surveys reveal Whites' views of Black Americans are changing recently, an important result given conventional wisdom that these are stable orientations. I test four possible explanations for these shifting views: genuine attitude change, social desirability, partisan expressive responding, and changing racial attitude measure performance. Importantly, these explanations produce observationally equivalent survey toplines. To adjudicate between them I use the measurement equivalence framework and examine how Whites answer the racial resentment measure. Evidence from multi-group confirmatory factor analysis models supports genuine attitude change. Substantively, this suggests these changes may have important political implications. Methodologically, it suggests partisan expressive responding may have limits, indicates social desirability pressures have not changed how Whites answer at least one racial attitude measure, and offers additional validity evidence for the racial resentment measure.

Replication Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/Y10EQV

Word Count: 9974

Surveys since the 1950s reveal marked changes in White Americans' racial attitudes. Whites increasingly endorse principles of racial equality–though often uncoupled from support for implementation–and oppose social distance (Schuman, Steeh, Bobo, and Krysan 1997; Krysan and Moberg 2016). Even their views of Black Americans, conventionally seen as stable orientations (Sears and Henry 2005; Tesler 2015), appear to be shifting of late (Hopkins and Washington 2020). As one stark example, American National Election Studies data reveal a substantial recent change in White Democrats' levels of racial resentment–their explanations for Black Americans' social and economic status (Kinder and Sanders 1996; Kam and Burge 2018). This group's average on this measure drops by 14 points, a 26% decrease, between 2012 and 2016 (Engelhardt 2019*b*).[1] Cast against conventional understandings these changes are arresting. Group evaluations tend to show substantial temporal stability (Tesler 2015). That these attitudes potently influence Whites' political thinking and behavior makes understanding why these shifts occurred critical (Hopkins 2019; Sides, Tesler and Vavreck 2018; Tesler 2016). Further, if they can change, then conventional wisdom about the status of these views relative to other orientations requires clarification (Tesler 2015).

[1] Similarly, the gap in White Democrats' average ratings of Black and White Americans on 101-point feeling thermometers is 3 points in 2016 (72 Black, 75 White), down from 9 points in 2012 (68 Black, 75 White). White Republicans exhibit little change (2012: 64 Black, 75 White; 2016: 63 Black, 73 White). In contrast, Hopkins and Washington (2020) report positive shifts in stereotype endorsement among Whites irrespective of party after 2016 using panel data.

I examine four possible explanations for these recent changes as a case to understand shifting racial attitudes more generally. Importantly, these theories of change produce observationally equivalent survey toplines, limiting extant investigations into these trends (e.g., Engelhardt 2020; Hopkins and Washington 2020). All offer plausible explanations for surveys finding Whites reporting more positive views of Black Americans. Understanding which matter is of paramount importance because they carry significantly different implications for how scholars interpret and study dynamics related to White racial attitudes and politics. Substantively, they differ in what consequences, if any, observed attitude change might have. Methodologically, they offer different perspectives about the nature and study of racial attitudes and public opinion.[2]

The first theory emphasizes attitude conversion. This theory, which I label the *genuine* explanation, places learning and, often, the contributions of external forces center stage. Positive outgroup exemplars (Goldman and Hopkins 2019; Goldman and Mutz 2014), norm clarification (Hopkins and Washington 2020), political elites' rhetoric and position-taking (Engelhardt 2020), or shifting social structures (Bobo 1999) can motivate individuals to change their racial attitudes.[3] New information or altered structural relations can lead some Whites to view racialized groups

[2] I focus throughout on explicit attitudes captured by survey self-reports. I do not investigate implicit attitudes because their conceptualization and measurement requires a related but distinct theoretical and analytical lens for evaluating dynamics (Greenwald and Lai 2020). I discuss this limitation, and descriptive analyses addressing it, in the conclusion.

[3] It is unlikely another potential mechanism–cohort replacement–explains these recent changes given their size and pace. This explanation seems better able to describe longer-term trends (Schuman et al. 1997).

in a different light. For instance, Engelhardt (2019*a*) reports that between 2011 and 2016, White survey respondents' views of Black Americans improved among those who reported watching MSNBC but the views of those watching Fox News or no similar programming did not change. That this and other recent work marshals evidence of individual-level change using panel data makes this argument plausible (e.g., Engelhardt 2020; Hopkins and Washington 2020), though unexpected given prevailing understandings about these attitudes' persistence after childhood formation (Sears and Henry 2005; Tesler 2015).

But even individual-level changes need not reflect genuine attitude conversion. One alternative theory holds that social norms affect Whites' responses to racial attitude measures, an explanation I label the *socially desirable* view. Scholars have long understood that perceptions of normatively appropriate views can lead people to misrepresent their attitudes (Schuman et al. 1997; Berinsky 1999; Crandall, Eshleman and O'Brien 2002; Hopkins 2009; Piston 2010; Stark, Maaren, Krosnick, and Sood 2019), with these motivations stemming from concerns both external and internal to the individual (Plant and Devine 1998; Paulhus 2002). While norms regarding the expression of prejudice politically have a long history (Mendelberg 2001), these prohibitions may be changing (Valentino, Neuner and Vandenbroek 2018). Group exemplars or important events–including the 2016 presidential election–can change beliefs about social norms (Crandall, Miller and White II 2018), making their consequences context-specific (Blinder, Ford and Ivarsflaten 2013; Hatchett and Schuman 1975; Hopkins 2009). Whites may now feel comfortable reporting attitudes that in past years were proscribed or they now restrict reporting their true beliefs due to increased scrutiny of negative outgroup attitudes. Their concern with how others evaluate them may have shifted and thus positive trends come from Whites deliberately mischaracterizing

their attitudes. That such external motivations may disproportionately shape prejudice's expression makes this potential change particularly important (Plant and Devine 1998).

A third position contends that racial attitude change, if diverging by partisanship, comes from identity performance. This *expressive* account springs from recent work suggesting individuals' party attachments may motivate them to use surveys and similar data collections to promote their party's position (Bullock, Gerber, Huber, and Hill 2015; Prior, Sood and Khanna 2015; Khanna and Sood 2018; Schaffner and Luks 2018). A desire to be the best partisan, and a member of the best team, leads people to willfully mischaracterize their true views. The same sources of identity-based misreporting of beliefs this literature at least implicitly draws from can also motivate attitude reports that conform to group norms (Ellemers, Spears and Doosje 2002).[4] From this perspective, White Democrats do not actually hold more positive attitudes about Black Americans. Instead, they know that the party line includes esteeming people of color. They then offer this position on survey questions to demonstrate they are good partisans, mischaracterizing

[4] Among several effects, group attachments motivate representing the group in the most positive light and provide information on how to think and act. Both can produce outcomes resembling expressive responding, though through distinct processes. One result comes from deriving esteem from identifying with the group. Positive esteem is harder to receive if the group is connected with unfavorable traits or positions. The second follows from identities simplifying a complex world. One way to arrive at an attitude is to identify what typical group members think or say and adopt this position. Changes in either the incentive to take a group typical position or esteem the group provides could yield survey results polarizing by party. Identifying which matters is beyond the scope of this work.

their underlying beliefs by adopting the party position. Positive, and polarizing, racial attitudes come from partisans editing their underlying beliefs.

The final theory, the *measurement* explanation, emphasizes varied attitude instrument quality. Scholars rely on socialization to explain racial attitudes' formation and persistence (Goldman and Hopkins 2020; Sears and Henry 2005). But changes in the information environment can shift how these socializing forces function, a possibility some accounts of racial attitude change directly incorporate when describing the shifting language of prejudice (Kinder and Sanders 1996; Sears and Henry 2005; see also Bobo 1999). The Obama administration, a diversifying country, and active social movements may have changed how race is discussed and understood (DeSante and Smith 2020*a*,*b*; Krysan and Moberg 2016). Relatedly, partisan differences in measure interpretation are possible because while political elites of all stripes discuss race (Gillion 2016), how they do varies markedly (Engelhardt 2019*a*). Racial attitude measures calibrated in another social and political context may thus be inappropriate for capturing these specific attitudes today because of changes in how Whites understand race. For example, consider two standard items measuring racial resentment (Kinder and Sanders 1996). Respondents are asked if they agree or disagree that "Generations of slavery and discrimination have created conditions that make it difficult for Blacks to work their way out of the lower class." Another item states "Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors." Whites may have divergent interpretations of these items because of changes in political context (DeSante and Smith 2020*a*,*b*) and/or how political elites talk about race (Engelhardt 2019*a*). Survey measures

can still indicate more positive attitudes, but what these attitudes are is less clear and inconsistent with prior opinion readings.

Using these post-2012 attitude changes as a motivating case, I test these explanations by using data from five iterations of the American National Election Study and to analyze the racial resentment measure within a multi-group confirmatory factor analysis framework. This technique assesses the relationship between observed measure responses and unobserved racial resentment. Doing so overcomes observationally equivalent survey toplines by instead testing whether the links between attitudes and survey measures vary in ways each theory predicts. Further, assessing racial resentment not only allows for evaluating attitude change explanations, its importance in work studying White racial attitudes' political relevance makes understanding sources of measure responses particularly consequential (Sides et al. 2018; Tesler 2016). My analyses suggest the *expressive* or *socially desirable* theories do little to explain these recent changes. Nor does much evidence support changing measure interpretation like the *measurement* position holds. The results point to the *genuine* view.

Racial Attitude Change and Measurement Equivalence

While these four explanations can produce the same observed racial attitude patterns, they differ in how they conceptualize the link between an attitude and its survey measure. For a given racial attitude measure, the *genuine* explanation says it captures the same attitude, to the same degree, for all Whites. The *expressive* and *socially desirable* views hold that while the measure captures the same attitude, different types of individuals over- or underreport their beliefs relative to their

true attitudes. Finally, the *measurement* position holds that the measure inconsistently captures the intended racial attitude because interpretation varies across types of people.

These observable implications align with levels of measurement equivalence in psychometrics. Used to offer evidence for the validity of group comparisons on a construct, measurement equivalence requires that the relationship between someone's unobserved attitude and responses to an attitude measure be independent of individual characteristics (Vandenberg and Lance 2000). Evidence supporting measurement equivalence is consistent with the *genuine* view that underlying attitudes have changed. In contrast, different equivalence violations offer evidence regarding the remaining three explanations.

The following regression equation shows mathematically how the measurement equivalence requirement can help adjudicate between these explanations.

$$y_{ijg} = \alpha_{jg} + \lambda_{jg}\xi_i + \varepsilon_{jg}$$

Individual *i* in group *g*'s response (*y*) to item *j* relates to her unobserved racial attitude (ξ) via an item-specific regression slope (λ), intercept (α), and error term (ε). For a measure to be equivalent–to capture the same construct, to the same degree, for all individuals–it must be consistently relevant, uniformly interpreted, and not incorporate concerns unrelated to latent racial attitude. One's score on an observed item *y* depends on ξ and nothing else.

Support for the *genuine* explanation comes from meeting these three requirements. For a given survey measure, measure items are consistently relevant if all items have a statistically significant relationship with ξ across groups. Regression slopes λ reliably differ from 0. The consistent meaning requirement holds that not only do these slopes differ from 0, the magnitude of a given item's slope does not vary across groups. The effect of a unit shift in ξ on y_j is the same

for all individuals. Further, the measure does not capture irrelevant considerations whose prevalence varies across groups, meaning each item's expected average response is consistent ($\alpha_{j,1} = \alpha_{j,2}$). On a given item, the regression parameters are the same for all individuals. Consequently, only variation in ξ explains variation in y.

But violations of measurement equivalence offer evidence for other explanations. Failing to meet the requirement that a measure does not also capture trait-irrelevant considerations supports the *expressive* and *socially desirable* positions. Each explanation argues concerns unrelated to racial attitudes–party reputation, social desirability–lead Whites to mischaracterize underlying beliefs. Support for these explanations manifests if intercepts, but not regression slopes, vary by individual type. Although people interpret the measure in the same way, these trait-irrelevant considerations shift the regression line relating ξ to y_j up or down (e.g., $\alpha_{j,1} + \xi > \alpha_{j,2} + \xi$; racial attitude in group 2 must be larger than racial attitude in group 1 to observe the same response on y_j).[5]

Evidence for the *measurement* explanation concerns the relevance and interpretation requirements. Changes in measure quality manifest as variation in the size of the relationship between items and racial attitude across groups ($\lambda_{j,1} \neq \lambda_{j,2}$) or as a special case where no measure relationship with racial attitude exists in one group ($\lambda_{j,1} = 0, \lambda_{j,2} > 0$). An interaction between item and individual characteristics exists, causing observed responses differences. Both results support the *measurement* explanation because the link between racial attitude and its measure reliably

[5] y may include ξ and considerations like social desirability but a measure remains equivalent. Error arises if this nuisance dimension's influence, or prevalence, varies across groups.

varies. The measure does worse at capturing attitude in one group compared to another, perhaps not measuring it at all.

[Table 1 about here]

Table 1 summarizes each explanation's observable implications within this framework. To illustrate, consider again White Democrats' changing views of Black Americans between 2012 and 2016. The individual characteristic of interest is survey year. Defining groups this way, the *genuine* explanation receives support if item regression slopes are the same magnitude ($\lambda_{j,2012} = \lambda_{j,2016}$) and Democrats do not systematically underreport negative views in 2016 compared to 2012 ($\alpha_{j,2012} = \alpha_{j,2016}$). Only change in latent group evaluations (ξ) can explain change in y_j between 2012 and 2016. Evidence for the *socially desirable* and *expressive* explanations manifests if regression slopes do not vary but Democrats systematically underreport negative views in 2016 compared to 2012 ($\alpha_{j,2012} > \alpha_{j,2016}$). 2016 respondents have to hold more negative views (ξ) to have the same observed response (y_j). Finally, if the *measurement* argument holds, then the relationship between latent group evaluations and their associated measure differs across groups. One or more items may have a significant relationship with the attitude in 2012 but not in 2016. Likewise, one or more item regression slopes differ in magnitude between 2012 and 2016 ($\lambda_{j,2012} \neq \lambda_{j,2016}$).

As this example illustrates, this framework still features an observational equivalence between the *socially desirable* and *expressive* explanations. They both explain differences in self-reports as distortions from individuals' underlying attitudes manifesting in intercept differences. Fortunately, however, these motivations concern different types of individuals. Consequently, to test these hypotheses I use the measurement equivalence framework and compare different

groups of White Americans where these explanations should most likely receive support. Consistency in findings across comparisons strengthens the case for an explanation.[6]

To test the *socially desirable* argument I use survey mode. If changes in social desirability concerns explain changes in racial attitudes, then intercept differences should manifest when comparing Whites completing face-to-face interviews with those completing the same measure online where these pressures are less salient ($\alpha_{j,f\,2f} < \alpha_{j,web}$). In-person respondents must score higher on latent racial attitude to offer the same observed item response as web respondents. Particularly helpful support for this position over the *expressive* explanation would come from within-party differences in responding across mode. While not dispositive, it is unlikely that identity-based response motivations vary by mode in ways complicating this comparison.

I take two approaches to test the *expressive* explanation. First, I look within partisan identifiers and compare responses across survey year. If White Democrats underreport negative

[6] Critically, this framework focuses on average differences across types of individuals and assumes a linear effect of latent attitude on observed responses. Consequently, it does not address whether individuals at different levels of latent attitude have different response motivations. The most prejudiced may have the greatest incentive to edit their attitudes. The present framework focuses on specific empirical trends in group averages. For instance, if the *socially desirable* argument explains observed changes, then a sufficient number of more prejudiced individuals must now be altering their self-reports. They have more to hide, but this motivation presumably varies in import by survey mode, with context relaxing pressures (Blinder et al. 2013; Hatchett and Schuman 1975; Stark et al. 2019). The present approach addresses this by focusing on types of individuals who have the most to gain from expressing certain attitudes, not individuals with a given trait level. Because this distinction between type of individual and trait level is important, I report analyses in OA5.4 (pgs 27-30) which offer some insight into potential variation by trait level and feature substantive results similar to those I report below.

racial attitudes in 2016, then intercept differences should manifest comparing this year to prior years (e.g., $\alpha_{j,2016} < \alpha_{j,2012} = \alpha_{j,2008} \ldots = \alpha_{j,g}$). 2016 interviewees would need to have higher values of latent racial attitude to offer the same observed item response as their copartisans interviewed in previous years. Second, I compare Democrats' and Republicans' responses within a given year. Again, if Democrats are underreporting their attitudes, then intercept differences should manifest across parties. They need to score higher on latent racial attitude to provide the same response to a survey item than Republicans.

All comparisons test the *measurement* view. Variation across mode, party, or year in the relationship between survey measure items and latent racial attitude indicates that changes in measurement explain observed trends, not belief editing or genuine attitude change.

I do not view any individual test as conclusive. I instead look for consistent results across places where evidence for each explanation should manifest to see where most support exists (cf. Berinsky 2018, who uses experiments to bound the extent of expressive survey responding). Consistent results offer stronger evidence for a given explanation.

Data and Method

I use data from American National Election Study surveys spanning 2000-2016. I focus on attitudes about Black Americans as measured by racial resentment (Kinder and Sanders 1996; Tarman and Sears 2005; Kam and Burge 2018). This measure captures explanations for Black people's social and economic status, a cultural manifestation of negative racial attitudes incorporating perceptions of norm violation. Table 2 contains the measure's question wording. Using racial resentment not only allows for testing attitude change explanations, the construct's importance in scholarship investigating the political relevance of Whites' racial attitudes makes

understanding sources of measure responses particularly consequential (Sides et al. 2018; Tesler 2016).

[Table 2 about here]

I analyze these data using multi-group confirmatory factor analysis and a well-validated approach to testing measurement equivalence (Brown 2015).[7] This procedure assesses changes in model fit between a series of nested models (Davidov 2009 and Pérez and Hetherington 2014 offer political science applications; Wicherts, Dolan and Hessen 2005 conduct an investigation like this paper). The first establishes whether all of the racial resentment measure's items have a significant relationship with latent racial resentment. I label this the equal form requirement. If this model exhibits poor fit, then the *measurement* explanation receives support because measure items inconsistently capture attitude. The second model determines whether the measure has consistent meaning across groups, which I call the equal factor loadings requirement. To do this, I constrain each item's factor loading between groups to test whether item regression slopes are the same. The *measurement* explanation receives support if this model fits worse than the equal form model. The final model tests whether Whites systematically over-/underreport racial resentment. I constrain item intercepts across groups to establish the equal intercepts requirement. If this model fits reliably worse than the equal factor loadings model,

[7] Analyses use R (v3.5.0) and lavaan (v0.5-23.1097) (Rosseel 2012).

then evidence supports the *expressive* or *socially desirable* explanations depending on the group compared.[8]

Conventional tests focus on a significant χ^2 difference between models. But many recommendations propose considering multiple model fit measures (Brown 2015). Consequently, I consider changes in the comparative fit index (CFI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA).[9] Because these measures lack consistent thresholds establishing equivalence, I use Jorgensen and colleagues' (2018) permutation method to generate empirical distributions for each with which I conduct hypothesis tests for change in model fit.[10] Consistency across fit statistics offers clearer insight into whether change in fit is reliable.

[8] In the measurement equivalence literature the equal form, equal factor loadings, and equal intercepts requirements are known as configural, metric, and scalar equivalence. I borrow the former terms from Brown (2015) to aid interpretability.

[9] CFI compares model performance relative to a null assuming no relationships among measure items exist. SRMR identifies the average difference between the measure's model-implied correlation matrix and its observed correlation matrix. RMSEA is a parsimony correction index indicating if the model fits reasonably well, contrasting χ^2's perfect fit test (Brown 2015).

[10] I save fit measures from the estimated MG-CFA model then permute indicators on the grouping variable (survey mode, survey year, party), assign to each the related row from the original data, and reestimate the MG-CFA model and save those fit statistics, doing this 2000 times. This produces a distribution of changes in model fit to which I compare the actual change to assess extremity. Routines implemented in semTools (v0.4-14) (Jorgensen, Pornprasertmanit, Schoemann, and Rosseel 2016).

I pair these statistical benchmarks with substantive criteria to identify support for each explanation. I use tests of small difference where the null is a small difference in model fit instead of the null of no difference the statistical criteria use (MacCallum, Browne and Cai 2006).[11] Rejecting this null provides the relevant explanation with greater support than rejecting the null of no difference because it indicates the change in fit is likely substantively meaningful. Similarly, I also consider overall model fit. Even if change in fit on a measure is significantly different from 0, the value may still indicate good model quality. If this occurs, then this second criterion suggests evidence for the relevant explanation is limited because models supporting this

[11] This procedure uses RMSEA to specify a difference in fit small enough to be negligible. This difference establishes a critical value from a non-central χ^2. If $\Delta\chi^2$ is less than this value, then the true difference in fit is likely small and unimportant. To provide consistent comparisons across studies and variation in test power I follow MacCallum et al. (2006) and use several RMSEA value-pairs indexing small changes in model fit at different levels of model quality. These are: .03-.05, .05-.06, .05-.07, .04-.07 (power mean: .83, median: .89, minimum: .29. 67% of tests have power ≥ .80). The pair .03-.05, for example, defines a decrease in RMSEA from .03-.05, but in a range of still excellent model equality (RMSEA ≤ .05). In contrast, .05-.07 is a similarly-sized change, but with RMSEA nearer .08 model quality is worse. Using several pairs allows me to understand the degree to which these tests can find changes in model fit of different magnitudes (distance between pairs) where these changes have different substantive consequences (value pair values).

explanation require estimating more parameters. With no single criterion dispositive, consistency across criteria indicate where most evidence points.

Combining statistical and substantive criteria addresses whether test evidence supports a given explanation for observed change. For instance, statistical evidence may not correspond with parameter differences large enough to account for empirical trends, something substantive criteria help address. Pairing these criteria also addresses variation in statistical power across tests. Greater statistical power allows for uncovering smaller parameter differences. But smaller discrepancies likely have muted practical consequences. Using MacCallum et al.'s (2006) power analysis recommendations, 20 of 26 model fit comparisons have power to detect a small-to-moderate difference in fit while model quality remains high of at least .80 (minimum: .64).[12] The statistical tests are therefore most likely to uncover inequivalence with small but meaningful substantive consequences, the implication of interest for these tests. Together the criteria speak to whether adding constraints produces changes in fit that are not only statistically significant but also substantively meaningful and capable of explaining much observed change.

Study 1: Social Desirability and Mode Effects

I begin testing the *socially desirable* explanation by using survey mode variation in the 2016 ANES. If this explanation holds, then face-to-face interviewees should underreport racial resentment compared to their online counterparts. Item intercepts should be reliably lower. If this doesn't occur, then the *genuine* explanation receives support.[13]

[12] *Measurement* explanation tests are most underpowered (5 of 6 underpowered), specifically Study 3 comparing Democrats' responses over time (4 of 6).

[13] While portions of the face-to-face interviews were completed via CASI, the interviewer

The first panel in Table 3 presents fit information for the models establishing each equivalence level for this assessment.[14] The rows report the results for each model. The columns report fit statistics and *p*-values from the permutation tests assessing change in fit after imposing a given constraint.[15] For CFI, values above .90 are adequate, with those above .95 excellent. SRMR values should fall below .08, with those nearer 0 ideal. RMSEA values below .10 are adequate, with .05 or lower excellent (Brown 2015).

The panel's first row offers evidence that the racial resentment measure meets equal form. Model fit is excellent (CFI = 1.000, SRMR = .001, RMSEA = .000). The second row suggests that constraining factor loadings leads to worse model fit, with these changes reliable (all $p \leq .01$). Even so, all statistic values indicate good-to-excellent fit (CFI = .998, SRMR = .022, RMSEA = .039). Likewise, small difference tests support negligible change (all $p > .90$). The third row incorporates the equal intercepts test. Like the equal factor loadings test, model fit reliably declines but overall fit remains excellent (CFI = .995, SRMR = .030, RMSEA = .047) and small difference tests support no meaningful change (all $p > .90$). While changes in fit suggest support for the *measurement* and *socially desirable* explanations, even the most restrictive and parsimonious model sees fit hardly compromised.

While a most likely place to identify response variation consistent with the *socially desirable* view, mode differences in the 2016 ANES yield little support for this explanation. Nor

asked the racial resentment measure.

[14] The online appendix includes model parameter estimates (pgs 1, 3, 6-7, 9, 14).

[15] I focus on CFI, SRMR, and RMSEA for parsimony because they offer clearer information on absolute and relative fit than χ^2. The online appendix includes $\Delta\chi^2$ (pgs 1, 3, 6-7, 9, 14).

does evidence consistent with the *measurement* argument clearly manifest. The *genuine* position therefore receives initial support.

[Table 3 about here]

Study 2: Social Desirability and Mode Effects within Party

Study 1 offered little evidence that the *socially desirable* view explains observed differences. But it also considered response differences for all Whites. It could be the case that social desirability matters, but not for everyone. With White Democrats' attitudes shifting in particular since 2012 (Engelhardt 2019*b*), partisanship may condition these pressures' relevance. To test this I conduct the same analysis as Study 1 but look within party. If Democrats are more susceptible to providing normatively appropriate responses, those answering face-to-face surveys should underreport racial resentment relative to their copartisans interviewed online. Lacking such pressures, it is unlikely Republicans' responses vary like this.

Table 3's second panel contains the results first for Democrats and then Republicans.[16] I focus first on Democrats. Model fit for the equal form and equal factor loadings tests offer little support for the *measurement* explanation. Nor is evidence consistent with the *socially desirable* explanation. Model fit remains great after constraining intercepts (CFI = .998, SRMR = .028, RMSEA = .030), despite two fit measures seeing reliable decreases (SRMR and RMSEA) with CFI close (p = .062). This suggests equal intercepts is met. The most parsimonious model sees little decline in model fit, and all fit measures are well away from suggesting poor model quality (CFI =

[16] Partisans include independent leaners.

.90, SRMR = .08, RMSEA = .10). Similarly, all small difference tests support negligible fit change (all $p > .90$). With model quality remaining great after constraining intercepts, the evidence suggests mode does not clearly condition White Democrats' responses, supporting the *genuine* view.

The results for Republicans offer similar insight. The results also offer little support for the *measurement* explanation. The panel's fourth and fifth rows indicate great fit, establishing equal form and equal factor loadings. Further, while constraining item intercepts produces a reliable change in CFI, with ΔRMSEA close ($p = .069$), model quality remains great according to each statistic (CFI = .992, SRMR = .030, RMSEA = .043). It could be that social desirability pressures shape Republicans' responses, but this is weak evidence. Supporting this, tests of small difference indicate negligible change (all $p > .90$). With Republicans' responses equivalent, the *genuine* explanation again receives support.

That interview context does not condition partisans' responses to the racial resentment measure offers additional evidence that the *socially desirable* explanation unlikely accounts for observed changes in White racial attitudes. Likewise, little evidence supports the *measurement* position. The *genuine* perspective again receives support.

Study 3: Expressive Responding and Temporal Effects

My first test of the *expressive* account looks within partisan groups and compares responses over time. Several approaches exist to do so. One option uses survey year as a grouping variable, simultaneously testing all years. While this offers an omnibus assessment, it is limited because inequivalence could come from relationships diverging between years apart from 2016, the key comparison year for evaluating attitude change explanations given recent trends (Engelhardt

2019*b*; Hopkins and Washington 2020). Instead, I iteratively compare the 2016 ANES face-to-face sample with the face-to-face interviews in each ANES including the racial resentment measure from 2000-2012, doing so separately for Democrats and Republicans. I can thus determine whether and how responses in 2016 differ from prior years, more clearly testing the *measurement* and *expressive* positions. Further, 2000 and 2004 serve as appropriate comparison years because political conflict in these years focused away from race, unlike years like 1988 or 1992 where racialized issues featured more prominently in campaign content (Hillygus and Shields 2008), potentially affecting responses. Inequivalence between 2000 and 2004 and later years should most likely stem from the *measurement* and *expressive* explanations, not features of these baseline years.

Panels 3 and 4 in Table 3 provide the results for Republicans and Democrats, respectively. In each the results first compare 2016 with 2000, while the remainder compare 2016 with 2004, 2008, and 2012. I take the results for Republicans first. All tests in panel 3 offer little evidence for the *measurement* argument. Each comparison supports the equal form and equal factor loadings requirements. Across comparisons, most fit measures indicate good, typically excellent, fit.[17] Further, no evidence clearly corroborates the *expressive* explanation. All comparisons establish the equal intercepts requirement with well-fitting models. In all tests involving the most parsimonious model the CFI is above .95 and SRMR and RMSEA below .05, indicating great model

[17] RMSEA for the 2000-2016 and 2004-2016 equal form tests is adequate (< .08). Adding constraints improves fit by adding degrees of freedom.

quality. Likewise, all small difference tests support substantively negligible change in fit (all $p >$ .75).

This evidence suggests claims that Donald Trump's election invigorated the expression of otherwise censored negative racial attitudes may require some qualification. It does not appear to be the case that Republicans avoided expressing negative view before 2016 as captured by this measure. Trump may have instead genuinely created negative attitudes (but see Hopkins and Washington 2020) or encouraged behaviors consistent with them (Schaffner 2020).

Table 3's fourth panel shows less consistent patterns for Democrats, but the *measurement* and *expressive* perspectives receive at best limited support. Consider rows 1-3, the results from the 2000-2016 comparison. Row 1 supports equal form with model fit excellent (CFI = 1.00, SRMR = .002, RMSEA = .000). Moving to the next row, model fit becomes mixed, with changes reliable. While CFI and SRMR remain good (.990 and .053), RMSEA is adequate (.075). Equal factor loadings may not be met. Based on statistical criteria, the *measurement* explanation receives some support. But tests of small difference suggest this may not be substantively meaningful (all $p > .21$). Change in fit is associated with at best minimal substantive consequences. The *measurement* perspective receives weak support.

Row 3 offers some support for the *expressive* view, but this is also weak. While CFI is excellent (.979), SRMR and RMSEA indicate at best adequate model fit (.072 and .088). This is initial support for the *expressive* view. But all three statistics support at least acceptable model quality, and small difference tests provide similar evidence (all $p > .38$). The results, while informative, offer limited support for the *expressive* position as the best explanation for observed patterns.

Panel 4's next three rows support measurement equivalence between 2004 and 2016. The *measurement* explanation receives little support with models establishing equal form (row 1) and equal factor loadings (row 2). Despite change in fit after imposing the equal factor loadings constraint, none are reliable and overall fit remains excellent (CFI = 1.000, SRMR = .030, RMSEA = .011). The results in row 3 also indicate good fit (CFI = .995, SRMR = .032, RMSEA = .047), despite statistically significant changes on two measures. Likewise, small difference tests suggest no meaningful change in fit (all $p > .70$). Despite some evidence from statistical criteria, the parsimonious equal intercepts model offers the most substantively meaningful characterization of Democrats' responses.

The results comparing 2008 and 2016 offer mixed evidence for the *expressive* position. The first two rows indicate models establishing equal form and factor loadings, again offering little support for the *measurement* explanation. Fit measures generally indicate good fit (RMSEA is at best adequate, though still acceptable). Row 9, however, suggests equal intercepts is not met. While CFI indicates good fit (.973), SRMR and RMSEA indicate adequate fit (.067 and .098), with all changes reliable. Statistical criteria support the *expressive* explanation and small difference tests, while inconclusive, are suggestive (all $p > .09$).

The last 3 rows offer consistent support for the *genuine* explanation comparing 2012 and 2016. Rows 10 and 11 support equal form and factor loadings, evidence against the *measurement* account. All three measures indicate excellent fit. The measure also meets equal intercepts, evidence inconsistent with the *expressive* view. All statistics again show excellent fit (CFI = .997, SRMR = .026, RMSEA = .031), with only ΔRMSEA reliably changing. The parsimonious model again best characterizes responses. Further, that the marked changes for Democrats occur between

2012 and 2016 makes this particularly meaningful support for the *genuine* argument over the *expressive* account.

Study 3 offers little consistent evidence for the *expressive* position. All comparisons among Republicans meet measurement equivalence, supporting the *genuine* explanation. Inconsistencies in measurement equivalence do arise among Democrats, the most likely place to find evidence for the *measurement* and *expressive* positions. But the results weakly and inconsistently support these explanations. Most evidence again supports the *genuine* account. This is reinforced by the fact that patterns are idiosyncratic across year comparisons, an unlikely outcome if expressive responding completely explains changes in Democrats' responses. Additional analyses using 2008 and 2012 as baselines instead of 2016 support this (OA, pgs 10-13). The evidence for mischaracterizing racial resentment by party is limited.

Study 4: Expressive Responding and Party Effects

I conclude my investigation by comparing the racial resentment measure's performance across parties. While Study 3 tested the *expressive* account using intraparty reactions to potential contextual changes, this assessment uses cross-party comparisons in 2016. If the *expressive* position holds, then Democrats' and Republicans' responses should differ. This should manifest, in particular, as Democrats underreporting racial resentment. They should need to score higher on latent racial resentment to provide the same observed response as a Republican if they edit their true attitudes.

Table 3's final panel presents the results using the 2016 ANES. To ensure that mode does not intersect with partisanship in shaping responses I conduct separate analyses on the face-to-face and web respondents.

I begin with the face-to-face group. Panel 5's first row indicates the measure meets equal form with all statistics indicating excellent model quality. The second row, however, offers evidence for the *measurement* position. While the CFI indicates excellent fit (.986), the SRMR and RMSEA are only adequate (.057 and .084), and changes on all are reliable or suggestive (ΔRMSEA p = .056). Generally, though, support seems limited as tests of small difference indicate negligible change in fit (all $p > .25$).

The results for the equal intercepts test offer some support for the *expressive* explanation. The CFI remains excellent (.970), and SRMR and RMSEA barely adequate (.079 and .097). Further, changes on the first two measures are reliable. But small differences tests all suggest negligible change (all $p > .28$), yielding mixed evidence. These results, and acceptable model quality, suggest that while the *expressive* perspective receives some support, it likely does not fully explain observed partisan differences.

The remaining results show similar patterns for web respondents. The measure meets equal form (row 4) but not equal factor loadings (row 5), offering some support for the *measurement* explanation. The CFI and SRMR support good fit for the second model (.990 and .055), and RMSEA only adequate (.076). Furthermore, changes across all measures are statistically significant. But small difference tests again suggest fit changes is not decisive (all $p > .17$). The *measurement* explanation receives some support.

Like the face-to-face respondents, row 6 supports the *expressive* account. Relative to the equal factor loadings model, fit for the equal intercepts model is reliably worse. The CFI remains excellent (.978), but the SRMR and RMSEA are only adequate (.074 and .091). Consistent with

expectations this is due to underestimation of racial resentment for Democrats relative to Republicans. But like the face-to-face group, substantive criteria indicate the *expressive* explanation appears unable to explain fully observed partisan differences in racial resentment levels. Fit change, per small difference tests, exists but remains negligible in size (all $p > .28$). The *expressive* explanation, while potentially important, appears incapable of accounting for group differences based on this test.

Contrasting Study 3, Study 4 offers more support for the *expressive* explanation. But if it explains recent changes, then party-based motivations for responding must have changed. These analyses offer a single party-based comparison potentially conflating differences from contextual change–the *expressive* view–with more durable divides in measure interpretation based on one's political orientation (Feldman and Huddy 2005, but see Enders 2019). These are important results but unrelated to the present investigation. To address whether results are unique to 2016, and therefore more supportive of the *measurement* and *expressive* accounts, I conducted additional analyses using the face-to-face respondents from the 2000, 2004, 2008, and 2012 ANES surveys (OA pgs 17-8). The results offer little evidence for either the *measurement* or *expressive* positions. The present analyses therefore appear to offer less support for the *expressive* explanation and instead suggest more consistent party-based differences.

Conclusion

That racial attitudes potently shape Whites' political thinking makes understanding reasons for change critically important. Even if carefully identified, observed trends may still feature observationally equivalent explanations. I compare four explanations used to explain changing

racial attitudes: genuine change, expressive change, socially desirable change, and measurement change. Using measurement models to evaluate responses to the racial resentment measure, I find evidence that *genuine* change appears best able to explain observed patterns. Evidence does manifest for the *expressive* and *measurement* positions, suggesting they may contribute some. Part of the difference in observed racial resentment between White Democrats and Republicans may come from varied measure interpretations. But support is inconsistent across tests and cannot fully explain observed shifts. Attitude change appears largely genuine.[18]

The results presented provide maximal possible evidence for explanations other than *genuine* change within this framework. The online appendix includes complementary information for each study suggesting these other explanations likely carry even less weight (pgs 1-17). First, inequivalence comes in degrees. Effect sizes for inequivalence violations (Gunn, Grimm and Edwards 2019) and tests for partial equivalence, a sufficient condition for measurement equivalence where the parameters (factor loadings, intercepts) for at least two items are equivalent across groups (Byrne, Shavelson and Muthen 1989), suggest the *measurement*, *socially desirable*, and *expressive* explanations cannot explain observed change. Second, in some instances inequivalence evidence runs opposite expectations. In Study 2, Republicans in the face-to-face sample if anything *overreport* racial resentment, evidence inconsistent with the *socially desirable* explanation (OA pgs 4-5). Likewise, in Study 4, error manifests among both Republicans

[18] Treating ordered items as continuous may affect the results. While an acceptable approach (Rhemtulla, Brosseau-Liard and Savalei 2012), analyses treating items as ordered offer similar insights. If anything the present results overstate inequivalence, providing the greatest possible, and still limited, evidence for explanations other than genuine change.

and Democrats (OA pgs 15-7). The measure better captures racial resentment among Democrats than Republicans while Democrats underreport racial resentment. While important insights into how partisans respond to this measure, the evidence is less conclusive regarding the contribution associated alternative explanations make to observed racial attitude change. They likely contribute some, but to a small degree.[19]

These results have important substantive and methodological implications. Substantively they support seeing changes in White racial attitudes as genuine (Hopkins and Washington 2020). Decreases in Democrats' racial resentment levels between 2012 and 2016 appear sincere (Engelhardt 2019*b*), not "cheap talk" reflecting a (racially) polarized political system (cf. Berinsky 2018; Bullock et al. 2015; Prior et al. 2015; Khanna and Sood 2018; Schaffner and Luks 2018). While attitude change may certainly be motivated by changing normative perceptions (Schuman et al. 1997; Crandall et al. 2002), the evidence here implies internalization beyond mere norm

[19] I also investigated the greatest extent to which each perspective may explain observed patterns (OA pgs 34-6). I calculated inequivalence effect sizes for the equal form and equal factor loadings models which assume that no loadings and no intercepts are equivalent, the greatest possible effect for the *measurement* and *socially desirable* or *expressive* explanations. This exercise suggests negligible *measurement* consequences, modest *socially desirable* effects, and more consequential effects for *expressive* in such best-case scenarios.

recognition. While certainly not decisive, the results suggest more than just an external motivation to avoid appearing prejudiced is at play (Plant and Devine 1998).[20]

These shifts imply broader behavioral and attitudinal implications. White Democrats may increasingly support policies addressing racial inequality and candidates championing the same. Consistent with this, coverage of 2020 Democratic primary voters suggests concerns about race and racial inequality featured in their candidate support decisions (Khalid 2019). Future work should systematically probe the depth of these commitments and their political effects to consider whether, and to what degree, these positive racial attitudes translate into support for restorative policies and behaviors reflecting improved intergroup beliefs. More generally, scholars should take seriously the potential contribution that positive racial attitudes play politically (see Chudy 2021; Engelhardt 2019*b*).

These results also offer several important methodological insights. First, they suggest that Whites' responses to racial attitude measures should be taken at face value, and that this holds for both Democrats and Republicans. While social scientists are rightly concerned with belief editing in surveys on sensitive topics like race or on questions connected with important

[20] I also investigated trends in reported internal and external motivations to respond without prejudice provided by participants in Harvard's Project Implicit. Between 2015 and 2018, when data are available, average external motivations change little while internal motivations strengthen (OA pgs 21-8). Nor do correlations between these orientations and feeling thermometer ratings of Black and White Americans change. While a unique sample, these results complement the evidence presented here.

individual identities, the present results do not suggest these concerns are necessarily problematic in all cases (see also Stark et al. 2019). Surveys should preserve space for explicit attitude measures. Further, consistencies in responding across survey mode suggest descriptive and inferential patterns found in online data collections are unlikely unique to mode, bolstering generalizability. With respect to expressive survey responding in particular, such processes may be more pernicious on fact-based items (Bullock et al. 2015; Schaffner and Luks 2018) than on attitude reports. Nor does party-motivated responding appear to manifest in ways suggesting such pressures respond to changes in the political context occurring over time. While certainly not ruling out expressive responding as a phenomenon, it may be isolated to specific survey item types.

Second, they offer necessary evidence for the racial resentment measure's comparability over time and across parties. Despite its formulation in a different social and political context, the measure still offers valid insight into the link between racial attitudes and politics despite apparent normative shifts (Valentino et al. 2018). Likewise, while some critique the measure for its inability to distinguish prejudice from principle on the political right (Feldman and Huddy 2005), Democrats and Republicans approach the measure in the same way (see also Enders 2019). While study 4's results do importantly suggest that the measure can be improved to better compare racial resentment across parties, its present form does not appear fatally flawed.

Finally, I highlight how approaches to establishing measurement equivalence can provide a framework for distinguishing between different observationally equivalence explanations for empirical phenomena (see also Wicherts, Dolan and Hessen 2005). While political scientists have considered measurement equivalence before (Davidov 2009; King, Murray, Salomon, and Tandon

2004; Pérez and Hetherington 2014; Pietryka and MacIntosh Forthcoming), these applications either introduce the problem or establish the equivalence of specific constructs rather than use these approaches to test substantive hypotheses. Theories of inequivalence can offer substantive insight.

Importantly, the insights and approach carry limitations. Paramount among them is the requirement that attitudes be captured by batteries, and the more items the better. Such a requirement is to the detriment of exhaustively assessing racial attitude measures to understand the various explanations I consider. It could be that while Whites' responses to the racial resentment measure do not exhibit inequivalence consistent with alternative explanations to genuine change, other measures may offer different evidence. This is plausible given racial animus's multi-dimensional nature (Kinder 2013). While limited, I offer evidence in the online appendix (pg 19-20) that a measure of attitudes about Muslim Americans introduced by Lajevardi (2020) also exhibits measurement equivalence by partisanship, suggesting limited expressive responding on the racial resentment measure is not unique.

Another limitation stems from studying self-reports. The preceding analyses do not engage with whether implicit attitudes have changed. If explicit attitudes shifted but implicit have not, then the often-divergent effects of explicit and implicit attitudes become increasingly consequential (Kalmoe and Piston 2013; Kinder and Ryan 2017). Further, the several explanations for observed attitude change still may matter. But if implicit have shifted, too, then this strengthens the preceding insights. Changes manifesting on indirect measures where response pressures associated with the *social desirability* and *expressive* accounts may carry less sway offer additional evidence against these positions. Likewise, with most measures capturing quick,

affective associations, the *measurement* explanation becomes less relevant (Greenwald and Lai 2020). Descriptive analyses of Project Implicit's Black-White Implicit Association Test reveal decreases in pro-White bias paralleling those found in self-reports (OA pgs 28-30), bolstering the present conclusions.

Understanding why attitudes change has important implications for interpreting and studying public opinion, especially regarding marginalized groups. This is a difficult task when myriad explanations offer observationally equivalent patterns. But by considering other observable implications, scholars can better understand observed trends, providing additional support for the arguments they develop through careful causal identification.

References

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43(4):1209–1230.

Berinsky, Adam J. 2018. "Telling the Truth about Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding." *The Journal of Politics* 80(1):211–224.

Blinder, Scott, Robert Ford and Elisabeth Ivarsflaten. 2013. "The Better Angels of Our Nature: How the Antiprejudice Norm Affects Policy and Party Preferences in Great Britain and Germany." *American Journal of Political Science* 46(2):841–857.

Bobo, Lawrence D. 1999. "Prejudice as Group Position: Microfoundations of a Sociological Approach to Racism and Race Relations." *Journal of Social Issues* 55(3):445–472.

Brown, Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research*. 2 ed. New York: Guilford Press.

Bullock, John G, Alan S Gerber, Gregory A Huber and Seth J Hill. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10(4):519–578.

Byrne, Barbara M, Richard J Shavelson and Bengt Muthen. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105(3):456–466.

Chudy, Jennifer. 2021. "Racial Sympathy and its Political Consequences." *The Journal of Politics* 83(1):122-136.

Crandall, Christian S, Amy Eshleman and Laurie O'Brien. 2002. "Social norms and the expression and suppression of prejudice: The struggle for internalization." *Journal of Personality and Social Psychology* 82(3):359–378.

Crandall, Christian S, Jason M Miller and Mark H White II. 2018. "Changing Norms Following the 2016 U.S. Presidential Election." *Social Psychological and Personality Science* 9(2):186– 192.

Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in a Comparative Perspective." *Political Analysis* 17(1):64–82.

DeSante, Christopher D and Candis Watts Smith. 2020*a*. "Less is More: A Cross-Generational Analysis of the Nature and Role of Racial Attitudes in the 21st Century." *Journal of Politics* 82(3):967–980.

DeSante, Christopher D and Candis Watts Smith. 2020*b*. *Racial Stasis*. Chicago: University of Chicago Press.

Ellemers, Naomi, Russell Spears and Bertjan Doosje. 2002. “Self and Social Identity*.” *Annual review of psychology* 53(1):161–186.

Enders, Adam M. 2019. “A Matter of Principle? On the Relationship Between Racial Resentment and Ideology.” *Political Behavior* .

Engelhardt, Andrew M. 2019*a*. “The Content of their Coverage: Contrasting Racially Conservative and Liberal Elite Rhetoric.” *Politics, Groups, and Identities* Online First https: //doi.org/10.1080/21565503.2019.1674672.

Engelhardt, Andrew M. 2019*b*. “Trumped by Race: Explanations for Race’s Influence on Whites’ Votes in 2016.” *Quarterly Journal of Political Science* 14(3):313–328.

Engelhardt, Andrew M. 2020. “Racial Attitudes through a Partisan Lens.” *British Journal of Political Science* First View https://doi.org/10.1017/S0007123419000437.

Feldman, Stanley and Leonie Huddy. 2005. “Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?” *American Journal of Political Science* 49(1):168–183.

Gillion, Daniel Q. 2016. *Governing with Words*. New York: Cambridge University Press.

Goldman, Seth K and Daniel J Hopkins. 2019. “When Can Exemplars Shape White Racial Attitudes? Evidence from the 2012 U.S. Presidential Campaign.” *International Journal of Public Opinion Research* 31(4):649–668.

Goldman, Seth K and Daniel J Hopkins. 2020. “Past Place, Present Prejudice: The Impact of Adolescent Racial Context on White Racial Attitudes.” *The Journal of Politics* 82(2):529–542.

Goldman, Seth K and Diana C Mutz. 2014. *The Obama Effect*. New York: Russell Sage Foundation.

Greenwald, Anthony G and Calvin K Lai. 2020. “Implicit Social Cognition.” *Annual review of psychology* 71(1):419–445.

Gunn, Heather J, Kevin J Grimm and Michael C Edwards. 2019. “Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes.” *Structural Equation Modeling: A Multidisciplinary Journal* .

Hatchett, Shirley and Howard Schuman. 1975. “White Respondents and Race-of-Interviewer Effects.” *Public Opinion Quarterly* 39(4):523–528.

Hillygus, D Sunshine and Todd G Shields. 2008. *The Persuadable Voter*. Wedge Issues in Presidential Campaigns Princeton: Princeton University Press.

Hopkins, Daniel J. 2009. “No More Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead about Black and Female Candidates.” *The Journal of Politics* 71(03):769–781.

Hopkins, Daniel J. 2019. “The Activation of Prejudice and Presidential Voting: Panel Evidence from the 2016 U.S. Election.” *Political Behavior*.

Hopkins, Daniel J and Samantha Washington. 2020. “The Rise of Trump, the Fall of Prejudice? Tracking White Americans’ Racial Attitudes 2008-2018 via a Panel Survey.” *Public Opinion Quarterly* 84(1):119–140.

Jorgensen, Terrence D, Benjamin A Kite, Po-Yi Chen and Stephen D Short. 2018. “Permutation Randomization Methods for Testing Measurement Equivalence and Detecting Differential Item Functioning in Multiple-Group Confirmatory Factor Analysis.” *Psychological Methods* 23(4):708–728.

Jorgensen, Terrence D, Sunthud Pornprasertmanit, Alexander M Schoemann and Yves Rosseel. 2016. *semTools: Useful tools for structural equation modeling*. R package version 0.4-14.

Kalmoe, Nathan P and Spencer Piston. 2013. “Is Implicit Prejudice against Blacks Politically Consequential? Evidence from the AMP.” *Public Opinion Quarterly* 77(1):305–322.

Kam, Cindy D and Camille D Burge. 2018. “Uncovering Reactions to the Racial Resentment Scale across the Racial Divide.” *The Journal of Politics* 80(1):314–320.

Khalid, Asma. 2019. “White Liberals Adopt More Progressive Positions On Race.” NPR.

Khanna, Kabir and Gaurav Sood. 2018. “Motivated Responding in Studies of Factual Learning.” *Political Behavior* 40(1):79–101.

Kinder, Donald R. 2013. Prejudice and Politics. In *The Oxford Handbook of Political Psychology*, ed. Leonie Huddy, David O Sears and Jack S Levy. New York: Oxford University Press pp. 812–851.

Kinder, Donald R and Lynn M Sanders. 1996. *Divided by Color*. Chicago: University of Chicago Press.

Kinder, Donald R and Timothy J Ryan. 2017. “Prejudice and Politics Re-Examined The Political Significance of Implicit Racial Bias.” *Political Science Research and Methods* 5(2):241–259.

King, Gary, Christopher J L Murray, Joshua A Salomon and Ajay Tandon. 2004. “Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research.” *American Political Science Review* 98(01):191–207.

Krysan, Maria and Sarah Moberg. 2016. “Trends in racial attitudes.” University of Illinois Institute of Government and Public Affairs.

Lajevardi, Nazita. 2020. *Outsiders at Home*. New York: Cambridge University Press.

MacCallum, Robert C, Michael W Browne and Li Cai. 2006. “Testing differences between nested covariance structure models: Power analysis and null hypotheses.” *Psychological Methods* 11(1):19–35.

Mendelberg, Tali. 2001. *The Race Card*. Princeton: Princeton University Press.

Paulhus, Delroy L. 2002. Socially Desirable Responding: The Evolution of a Construct. In *The Role of Constructs in Psychological and Educational Measurement*. Routledge pp. 49–72.

Pérez, Efrén O and Marc J Hetherington. 2014. “Authoritarianism in Black and White: Testing the Cross-Racial Validity of the Child Rearing Scale.” *Political Analysis* 22(3):398–412.

Pietryka, Matthew T. and Randall C MacIntosh. Forthcoming. “ANES Scales Often Don’t Measure What You Think They Measure – An ERPC2016 Analysis.” *Journal of Politics* https://pie-try-ka.com/publication/ap-anes-preregistration/AP%20-% 20ANES%20Preregistration.pdf.

Piston, Spencer. 2010. “How Explicit Racial Prejudice Hurt Obama in the 2008 Election.” *Political Behavior* 32(4):431–451.

Plant, E Ashby and Patricia G Devine. 1998. “Internal and external motivation to respond without prejudice.” *Journal of personality and social psychology* 75(3):811–832.

Prior, Markus, Gaurav Sood and Kabir Khanna. 2015. “You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions.” *Quarterly Journal of Political Science* 10(4):489–518.

Rhemtulla, Mijke, Patricia É Brosseau-Liard and Victoria Savalei. 2012. “When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions.” *Psychological Methods* 17(3):354–373.

Rosseel, Yves. 2012. “lavaan: An R Package for Structural Equation Modeling.” *Journal of Statistical Software, Articles* 48(2):1–36.

Schaffner, Brian F. 2020. *The Acceptance and Expression of Prejudice during the Trump Era*. New York: Cambridge University Press.

Schaffner, Brian F and Samantha Luks. 2018. “Misinformation or Expressive Responding?” *Public Opinion Quarterly* 82(1):135–147.

Schuman, Howard, Charlotte Steeh, Lawrence D Bobo and Maria Krysan. 1997. *Racial Attitudes in America: Trends and Interpretations*. 2 ed. Cambridge: Harvard University Press.

Sears, David O and P J Henry. 2005. Over Thirty Years Later: A Contemporary Look at Symbolic Racism. In *Advances in Experimental Social Psychology*, ed. Mark P Zanna. San Diego: Elsevier pp. 95–150.

Sides, John, Michael Tesler and Lynn Vavreck. 2018. *Identity Crisis*. Princeton: Princeton University Press.

Stark, Tobias H, Floor M van Maaren, Jon A Krosnick and Gaurav Sood. 2019. "The Impact of Social Desirability Pressures on Whites' Endorsement of Racial Stereotypes: A Comparison Between Oral and ACASI Reports in a National Survey." *Sociological Methods & Research* 1:004912411987595–27.

Tarman, Christopher and David O Sears. 2005. "The Conceptualization and Measurement of Symbolic Racism." *The Journal of Politics* 67(3):731–761.

Tesler, Michael. 2015. "Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed." *American Journal of Political Science* 59(4):806–824.

Tesler, Michael. 2016. *Post-Racial or Most-Racial?* Chicago: University of Chicago Press.

Valentino, Nicholas A, Fabian G Neuner and L Matthew Vandenbroek. 2018. "The Changing Norms of Racial Political Rhetoric and the End of Racial Priming." *The Journal of Politics* 80(3):757–771.

Vandenberg, Robert J and Charles E Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3(1):4–70.

Wicherts, Jelte M, Conor V Dolan and David J Hessen. 2005. "Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance." *Journal of Personality and Social Psychology* 89(5):696–716.

Table 1: Hypotheses and Implications

Explanation	Expectation	Testable Implication	
Genuine	Racial attitude measures capture the same construct and only that construct	$\lambda_{j,1} = \lambda_{j,2}$ $\alpha_{j,1} = \alpha_{j,2}$	Equal Slopes, Equal Intercepts
Socially Desirable	Racial attitude measures capture the same construct but not only that construct	$\lambda_{j,1} = \lambda_{j,2}$ $\alpha_{j,1} \neq \alpha_{j,2}$	Equal Slopes, Unequal Intercepts
Expressive	Racial attitude measures capture the same construct but not only that construct	$\lambda_{j,1} = \lambda_{j,2}$ $\alpha_{j,1} \neq \alpha_{j,2}$	Equal Slopes, Unequal Intercepts
Measurement	Racial attitude measures do not capture the same construct	$\lambda_{j,1} \neq \lambda_{j,2}$	Unequal Slopes

Table 2: Question Wording

Item	Question Wording
Deserve Less	Over the past few years, Blacks have gotten less than they deserve.
Past Discrimination	Generations of slavery and discrimination have created conditions that make it difficult for Blacks to work their way out of the lower class.
Special Favors	Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. (R)
Try Hard	It's really a matter of some people not trying hard enough; if Blacks would only try harder they could be just as well off as Whites. (R)

Note: Items marked (R) are reverse coded. Responses recorded on 5-point strongly agree–strongly disagree scales.

Table 3: Model Fit Statistics from Measurement Equivalence Tests

	Model	CFI	*p*-value	SRMR	*p*-value	RMSEA	*p*-value
Study 1: Mode	Equal Form	1.000		.001		.000	
	+ Equal Loadings	.998	.006	.022	.010	.039	.002
	+ Equal Intercepts	.995	.002	.030	.008	.047	.114
Study 2: Mode within Party							
Democrats							
	Equal Form	1.000		.002		.000	
	+ Equal Loadings	1.000	.864	.018	.294	.000	.815
	+ Equal Intercepts	.998	.062	.028	.006	.030	.028
Republicans	Equal Form	.999		.004		.032	
	+ Equal Loadings	.998	.353	.021	.404	.025	.824
	+ Equal Intercepts	.992	.018	.030	.100	.043	.069
Study 3: Year within Party							
Republicans 2000-2016							
	Equal Form	.993		.013		.072	
	+ Equal Loadings	.998	.993	.021	.904	.026	.982
	+ Equal Intercepts	.993	.120	.032	.198	.037	.152
2004-2016	Equal Form	.992		.013		.079	
	+ Equal Loadings	.994	.976	.024	.854	.045	.979
	+ Equal Intercepts	.992	.188	.035	.170	.040	.728
2008-2016	Equal Form	.997		.009		.049	
	+ Equal Loadings	.992	.162	.039	.125	.051	.138
	+ Equal Intercepts	.989	.194	.049	.118	.047	.382
2012-2016	Equal Form	.996		.010		.052	
	+ Equal Loadings	.992	.244	.030	.351	.047	.155
	+ Equal Intercepts	.991	.314	.039	.168	.040	.358
Democrats 2000-2016	Equal Form	1.000		.002		.000	
	+ Equal Loadings	.990	.013	.053	.006	.075	.009
	+ Equal Intercepts	.979	.005	.072	.016	.088	.128
2004-2016	Equal Form	1.000		.004		.000	
	+ Equal Loadings	1.000	.314	.030	.197	.012	.212
	+ Equal Intercepts	.995	.035	.032	.550	.047	.034
2008-2016	Equal Form	.997		.009		.071	
	+ Equal Loadings	.994	.108	.035	.083	.061	.196
	+ Equal Intercepts	.973	< .001	.067	< .001	.098	.018
2012-2016	Equal Form	1.000		.003		.000	

	+ Equal Loadings	1.000	.884	.014	.740	.000	.846
	+ Equal Intercepts	.997	.092	.026	.080	.031	.049
Study 4: Party							
Face-to-Face							
	Equal Form	.998		.008		.045	
	+ Equal Loadings	.986	.007	.057	.005	.084	.056
	+ Equal Intercepts	.970	.002	.079	.004	.097	.138
Web	Equal Form	1.000		.001		.000	
	+ Equal Loadings	.990	.001	.055	< .001	.077	< .001
	+ Equal Intercepts	.978	< .001	.074	< .001	.091	.074

Note: minima for acceptable fit are: .900 (CFI), .080 (SRMR), .100 (RMSEA). Reported *p*-values address change for the statistic after including constraint.

Online Appendix
Observational Equivalence in Explaining Attitude Change: Have White Racial Attitudes Genuinely Changed?

Contents

A Study 1: Social Desirability and Mode Effects

A.1 Main Text Models

Table A.1: Mode Equivalence

	Equal Form		Equal Factor Loadings		Equal Intercepts	
	Web	Face-to-Face	Web	Face-to-Face	Web	Face-to-Face
Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—
Try Hard	0.751	0.867	0.844	0.844	0.843	0.843
	(0.050)	(0.028)	(0.024)	(0.024)	(0.024)	(0.024)
Special Favors	0.748	0.945	0.900	0.900	0.898	0.898
	(0.046)	(0.028)	(0.024)	(0.024)	(0.024)	(0.024)
Past Discrimination	0.923	1.071	1.040	1.040	1.039	1.039
	(0.052)	(0.029)	(0.025)	(0.025)	(0.025)	(0.025)
Intercept Deserve Less	3.418	3.441	3.418	3.441	3.412	3.412
	(0.050)	(0.029)	(0.048)	(0.029)	(0.045)	(0.045)
Intercept Try Hard	3.143	3.077	3.143	3.077	3.070	3.070
	(0.053)	(0.030)	(0.053)	(0.030)	(0.041)	(0.041)
Intercept Special Favors	3.651	3.567	3.650	3.567	3.564	3.564
	(0.050)	(0.031)	(0.052)	(0.030)	(0.043)	(0.043)
Intercept Past	3.051	3.196	3.051	3.196	3.139	3.139
Discrimination	(0.054)	(0.032)	(0.055)	(0.032)	(0.048)	(0.048)
$\chi 2$	2		15		31	
DF	2		5		8	
CFI	1.000		0.998		0.995	
SRMR	0.001		0.022		0.030	
RMSEA [90% CI]	0 [0, 0.051]		0.039 [0.018, 0.063]		0.047 [0.031, 0.065]	
N	716	1912	716	1912	716	1912

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

A.2 Partial Equivalence and Substantive Effects

The main results use statistical criteria and parsimony to assess model performance. To this I add results from analyses using substantive criteria to assess model performance. I focus on two things: partial equivalence and effect size measures. Partial equivalence is a sufficient condition for measurement equivalence where the parameters (i.e., loadings, intercepts) for at least two items are equivalent across groups (Byrne, Shavelson and Muthen 1989). Measures remain equivalent if individuals share interpretations of, and responses to, some, but not all, items. Meeting partial equivalence, especially if inequivalence patterns are idiosyncratic, suggests alternative explanations have weaker support than the main results indicate. Effect sizes suggest how much item

inequivalence contributes to differences in predicted item scores between each group, given model parameter estimates (Gunn, Grimm and Edwards 2019).

Table A.2: Measurement Equivalence of Racial Resentment by Mode

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
Equal Form	1.530	1.000	0.001	0.000								
Equal Factor Loadings	15.200	0.998	0.022	0.039	13.700	0.006	-0.002	0.006	0.021	0.010	0.039	0.003
Equal Factor Loadings[1]	8.450	0.999	0.014	0.029	6.920	0.036	-0.001	0.027	0.013	0.077	0.029	0.019
Equal Intercepts[1]	24.700	0.996	0.020	0.044	16.300	0.002	-0.003	0.002	0.006	0.052	0.015	0.066
Equal Intercepts[1,2]	13.000	0.999	0.019	0.030	4.580	0.114	-0.001	0.079	0.004	0.080	0.001	0.186

Note: Models use *deserve less* to define the dimension. One error covariance estimated between *try hard* and *special favors*.
1: frees *special favors* loading; 2: frees *past discrimination* intercept

Table A.2 includes model fit information for each equivalence level test. Rows 1 and 2 mirror the main text results. But instead of proceeding to test equal intercepts, I explore if the equal factor loadings model meets partial equivalence. I first examine model modification indices. These measures offer an approximate change in model fit after unconstraining parameters (Brown 2015).[1] They suggest freeing *special favors* (modification index [MI]= 6.87, p = .029). Its factor loading is lower in the face-to-face group than the web group (expected parameter change [EPC]= -.163 and .044).[2] Row 3 shows that fit improves, but 3 measures still show reliably worse fit. Even so, no modification indices point to item loadings as contributing to worse model fit, results suggesting an appropriate model (Bollen 1989). [3] Freeing one item loading establishes partial equivalence. The *measurement* explanation receives little support.

Effect size measures and corrected group mean comparisons offer additional evidence against the *socially desirable* explanation. I use the SDI_2 and UDI_2 measures Gunn, Grimm and Edwards (2019)

[1] To control Type I error modification index p-values are calculated via Tukey's honest significant difference method, with modification index values determined within the permutation framework (Jorgensen et al. 2018).

[2] EPCs indicate how much a parameter may change if freely estimated in subsequent analyses. The EPCs of -.163 and .044 indicate that relative to an estimated loading of .900 in the constrained model, the factor loading for *special favors* in the face-to-face and web groups are approximately .737 and .944, respectively.

[3] The largest MI is 2.97 (*past discrimination*). But p=.148, offering little evidence fit will reliably improve.

introduce. UDI_2 captures how much inequivalence is present across the distribution of latent racial attitude. SDI_2 isolates this to variation in observed scores for the focal group (here, web respondents). Like Cohen's *d*, larger values denote greater practical consequences.[4] In no case do meaningful practical consequences manifest. In the partially equivalent equal loadings model, SDI_2 and UDI_2 estimates for *special favors* are .063 and .088. This holds in the partially equivalent equal intercepts model, with SDI_2 and UDI_2 estimates for *special favors* and *past discrimination* well below .20 indicating negligible practical effects (*special favors*: SDI_2 = .002, UDI_2 = .071, *past discrimination*: SDI_2 =−.114, UDI_2 = .114).

Inequivalence also does not alter insights into observed differences in racial resentment levels by mode. Following Nye and Drasgow (2011), I decompose observed mean differences into the part from genuine group differences (the psychometric feature known as impact) and that attributable to bias from partial inequivalence. This takes observed racial resentment scores, subtracts from them predicted scores using the parameters from the final estimated model establishing equal intercepts, and then takes the difference for this quantity across the two groups (i.e., face-to-face − web). The observed mean difference is -.004 points on the 5-point scale. But within this exists offsetting influences of bias (-.171, Cohen's *d* = -.178) and impact (.167, *d* = .175). Genuine group differences exist where face-to-face respondents report higher levels of racial resentment but where differences in measure performance negate this difference. Not only substantively small, this also runs against expectations from the *socially desirable* explanation where face-to-face respondents should score *lower* on racial resentment.

B Study 2: Social Desirability and Mode Effects within Party

[4] Gunn, Grimm and Edwards (2019) do not report effect size benchmarks for UDI_2 and SDI_2, so I use Cohen's *d* benchmarks as suggestive for narrative rather than definitive. Comparisons across models are more instructive. Critically, this also provides a more restrictive comparison than the .4, .6, .8 of small, medium, and large effects proposed for some equivalence measures (Nye et al. 2019), which are closely related to those I use here (Gunn, Grimm and Edwards 2019).

B.1 Main Text Models

Table B.1: Mode Equivalence by Party

	Democrats						Republicans					
	Equal Form		Equal Factor Loadings		Equal Intercepts		Equal Form		Equal Factor Loadings		Equal Intercepts	
	Web	Face-to-Face	Web	Face-to-Face	Web	Face-to-Face	Web	Face-to-Face	Web	Face-to-Face	Web	Face-to-Face
Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—	—	—	—	—	—	—
Try Hard	0.722	0.843	0.813	0.813	0.812	0.812	0.671	0.780	0.761	0.761	0.758	0.758
	(0.072)	(0.044)	(0.038)	(0.038)	(0.038)	(0.038)	(0.109)	(0.059)	(0.051)	(0.051)	(0.051)	(0.051)
Special Favors	0.847	1.005	0.964	0.964	0.962	0.962	0.517	0.697	0.654	0.654	0.651	0.651
	(0.073)	(0.047)	(0.040)	(0.040)	(0.040)	(0.040)	(0.081)	(0.051)	(0.044)	(0.044)	(0.044)	(0.044)
Past Discrimination	0.945	1.053	1.026	1.026	1.026	1.026	0.915	1.098	1.060	1.060	1.063	1.063
	(0.071)	(0.045)	(0.038)	(0.038)	(0.038)	(0.038)	(0.132)	(0.071)	(0.062)	(0.062)	(0.063)	(0.063)
Intercept Deserve Less	2.737	2.792	2.737	2.792	2.758	2.758	3.895	3.937	3.895	3.937	3.870	3.870
	(0.081)	(0.044)	(0.079)	(0.044)	(0.075)	(0.075)	(0.057)	(0.035)	(0.056)	(0.036)	(0.051)	(0.051)
Intercept Try Hard	2.442	2.399	2.442	2.399	2.386	2.386	3.627	3.585	3.627	3.585	3.554	3.554
	(0.085)	(0.045)	(0.087)	(0.045)	(0.066)	(0.066)	(0.061)	(0.038)	(0.061)	(0.038)	(0.045)	(0.045)
Intercept Special Favors	2.940	2.790	2.940	2.790	2.803	2.803	4.167	4.145	4.167	4.145	4.115	4.115
	(0.088)	(0.050)	(0.090)	(0.050)	(0.076)	(0.076)	(0.051)	(0.033)	(0.052)	(0.033)	(0.038)	(0.038)
Intercept Past Discrimination	2.421	2.505	2.421	2.505	2.462	2.462	3.492	3.724	3.492	3.724	3.611	3.611
	(0.084)	(0.048)	(0.084)	(0.048)	(0.077)	(0.077)	(0.068)	(0.041)	(0.068)	(0.041)	(0.057)	(0.057)
χ2	1		4		12		3		7		17	
DF	2		5		8		2		5		8	
CFI	1		1		0.998		0.999		0.998		0.992	
SRMR	0.002		0.018		0.028		0.004		0.021		0.030	
RMSEA [90% CI]	0 [0, 0.064]		0 [0, 0.056]		0.030 [0, 0.064]		0.032 [0, 0.091]		0.025 [0, 0.064]		0.043 [0.014, 0.071]	
N	275	762	275	762	275	762	372	899	372	899	372	899

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

B.2 Partial Equivalence and Substantive Effects

Table B.2: Measurement Equivalence of Racial Resentment by Mode within Party

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
Democrats												
Equal Form	0.695	1.000	0.002	0.000								
Equal Factor Loadings	4.190	1.000	0.018	0.000	3.500	0.362	0.000	0.865	0.016	0.294	0.000	0.815
Equal Intercepts	11.800	0.998	0.028	0.030	7.580	0.063	-0.002	0.062	0.011	0.007	0.030	0.028
Equal Intercepts[1]	5.820	1.000	0.021	0.000	1.630	0.436	0.000	0.768	0.004	0.147	0.000	0.717
Republicans												
Equal Form	3.300	0.999	0.004	0.032								
Equal Factor Loadings	6.950	0.998	0.021	0.025	3.650	0.416	-0.001	0.353	0.017	0.404	-0.007	0.824
Equal Intercepts	17.400	0.992	0.030	0.043	10.500	0.024	-0.006	0.018	0.008	0.100	0.018	0.069
Equal Intercepts[2]	8.190	0.999	0.025	0.016	1.240	0.553	0.001	0.822	0.004	0.298	-0.008	0.762

Note: Models use *deserve less* to define the dimension. One error covariance estimated between *try hard* and *special favors*.

1: frees *special favors* intercept; 2: frees *past discrimination* intercept

Table B.2 extends the results from Study 2. The top and bottoms panels compare Democrats' and Republicans' responses by mode, respectively. I focus first on whether the racial resentment measure

meets partial equivalence on the equal intercepts test in each case. For Democrats, modification indices suggest potential misspecification due to *special favors* (MI = 5.89, p = .060). Those completing face-to-face interviews appear to underreport racial resentment compared to those in the web sample as indicate by a likely higher estimated intercept (Expected parameter change $[EPC]_{f2f}$ = .137, EPC_{web} = -.035). This evidence is consistent with the *socially desirable* account. But the modification index value and associated p-value for change in fit do not suggest large fit improvements from freeing this parameter. But with this constraint potentially creating worse fit, and to be as generous as possible in identifying potential evidence for alternative explanations, I free the item intercept. As the top panel's final row shows, doing so produces a well-fitting model establishing the equal intercepts requirement. For Republicans, *past discrimination* appears problematic (MI = 9.21, p = .010). Face-to-face respondents may underreport racial resentment on this item (EPC = -.119) compared to the web group (EPC = .034). After freeing this intercept, the results indicate a well-fitting equal intercepts test (bottom panel, row 4). Support for the *socially desirable* view is limited.

Substantively, the inequivalence patterns identified in the main text suggest the *socially desirable* argument cannot explain observed differences.[5] The SDI_2 and UDI_2 effect size measures for Republicans are -.173 and .173, where the contrast in signs signifies issues in the face-to-face group. For Democrats, the effect size measures both equal .116. Further, effects on mean comparisons are inconsistent with the *socially desirable* explanation. For Democrats, the observed difference is .007 points, with measure bias (.170, Cohen's d = .174) cancelling out impact in the opposite direction (-.163, d = -.168). For Republicans, the observed gap is -.055 points, with this due to measure bias (-.181, d = -.219) also negating impact (.126, d = .155). While the results for Democrats support the *socially desirable* view, the opposite holds for Republicans. These differences are also substantively small with, if anything, larger practical effects among Republicans, a result inconsistent with the *socially desirable* explanation.

[5] For an explanation of these metrics, see Appendix section A.2

C Study 3: Expressive Responding and Temporal Effects

C.1 Main Text Models

Table C.1: Temporal Equivalence 2000 and 2004, Republicans

	2000							2004					
	Equal Form		Equal Factor Loadings		Equal Intercepts			Equal Form		Equal Factor Loadings		Equal Intercepts	
	2000	2016	2000	2016	2000	2016		2004	2016	2004	2016	2004	2016
Past Discrimination	1.000	1.000	1.000	1.000	1.000	1.000	Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—		—	—	—	—	—	—
Deserve Less	0.962	1.093	1.039	1.039	1.053	1.053	Past Discrimination	1.168	0.915	1.049	1.049	1.045	1.045
	(0.183)	(0.158)	(0.118)	(0.118)	(0.121)	(0.121)		(0.150)	(0.132)	(0.099)	(0.099)	(0.100)	(0.100)
Try Hard	0.617	0.734	0.690	0.690	0.695	0.695	Try Hard	0.717	0.671	0.704	0.704	0.694	0.694
	(0.119)	(0.103)	(0.078)	(0.078)	(0.078)	(0.078)		(0.096)	(0.109)	(0.071)	(0.071)	(0.071)	(0.071)
Special Favors	0.445	0.565	0.516	0.516	0.514	0.514	Special Favors	0.518	0.517	0.515	0.515	0.508	0.508
	(0.095)	(0.088)	(0.065)	(0.065)	(0.065)	(0.065)		(0.084)	(0.081)	(0.058)	(0.058)	(0.058)	(0.058)
Intercept Past	3.491	3.492	3.491	3.492	3.438	3.438	Intercept Past	3.387	3.492	3.387	3.492	3.397	3.397
Discrimination	(0.087)	(0.068)	(0.086)	(0.068)	(0.072)	(0.072)	Discrimination	(0.064)	(0.068)	(0.063)	(0.069)	(0.057)	(0.057)
Intercept Deserve	3.746	3.895	3.746	3.895	3.783	3.783	Intercept Deserve	3.777	3.895	3.777	3.895	3.794	3.794
Less	(0.073)	(0.057)	(0.073)	(0.057)	(0.069)	(0.069)	Less	(0.052)	(0.057)	(0.052)	(0.057)	(0.050)	(0.050)
Intercept Try Hard	3.548	3.627	3.548	3.627	3.558	3.558	Intercept Try Hard	3.690	3.627	3.690	3.627	3.636	3.636
	(0.071)	(0.061)	(0.071)	(0.060)	(0.057)	(0.057)		(0.055)	(0.061)	(0.055)	(0.060)	(0.047)	(0.047)
Intercept Special	4.237	4.167	4.237	4.167	4.167	4.167	Intercept Special	4.205	4.167	4.205	4.167	4.169	4.169
Favors	(0.062)	(0.051)	(0.062)	(0.050)	(0.047)	(0.047)	Favors	(0.046)	(0.051)	(0.046)	(0.050)	(0.038)	(0.038)
$\chi 2$	5		6		11			7		9		13	
DF	2		5		8			2		5		8	
CFI	0.993		0.998		0.993			0.992		0.994		0.992	
SRMR	0.013		0.022		0.032			0.013		0.024		0.035	
RMSEA [90% CI]	0.072 [0, 0.152]		0.026 [0, 0.088]		0.037 [0, 0.082]			0.079 [0.017, 0.148]		0.045 [0, 0.093]		0.040 [0, 0.079]	
N	228	372	228	372	228	372		377	372	377	372	377	372

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

Table C.2: Temporal Equivalence 2008 and 2012, Republicans

	2008						2012					
	Equal Form		Equal Factor Loadings		Equal Intercepts		Equal Form		Equal Factor Loadings		Equal Intercepts	
	2008	2016	2008	2016	2008	2016	2012	2016	2012	2016	2012	2016
Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—	—	—	—	—	—	—
Try Hard	0.428	0.671	0.526	0.526	0.538	0.538	0.926	0.671	0.798	0.798	0.786	0.786
	(0.080)	(0.109)	(0.065)	(0.065)	(0.065)	(0.065)	(0.126)	(0.109)	(0.084)	(0.084)	(0.084)	(0.084)
Special Favors	0.444	0.517	0.471	0.471	0.476	0.476	0.566	0.517	0.548	0.548	0.544	0.544
	(0.067)	(0.081)	(0.052)	(0.052)	(0.051)	(0.051)	(0.098)	(0.081)	(0.063)	(0.063)	(0.063)	(0.063)
Past Discrimination	0.991	0.915	0.952	0.952	0.965	0.965	1.371	0.915	1.130	1.130	1.109	1.109
	(0.134)	(0.132)	(0.095)	(0.095)	(0.094)	(0.094)	(0.182)	(0.132)	(0.112)	(0.112)	(0.111)	(0.111)
Intercept Deserve Less	3.981	3.895	3.981	3.895	3.989	3.989	4.062	3.895	4.062	3.895	4.033	4.033
	(0.046)	(0.057)	(0.046)	(0.057)	(0.045)	(0.045)	(0.050)	(0.057)	(0.051)	(0.056)	(0.048)	(0.048)
Intercept Try Hard	3.820	3.627	3.820	3.626	3.762	3.762	3.650	3.627	3.649	3.627	3.676	3.676
	(0.050)	(0.061)	(0.051)	(0.059)	(0.042)	(0.042)	(0.058)	(0.061)	(0.057)	(0.061)	(0.048)	(0.048)
Intercept Special Favors	4.211	4.167	4.211	4.167	4.213	4.213	4.222	4.167	4.222	4.167	4.221	4.221
	(0.042)	(0.051)	(0.042)	(0.050)	(0.035)	(0.035)	(0.049)	(0.051)	(0.049)	(0.050)	(0.039)	(0.039)
Intercept Past	3.619	3.492	3.619	3.492	3.610	3.610	3.524	3.492	3.524	3.492	3.561	3.561
Discrimination	(0.059)	(0.068)	(0.058)	(0.069)	(0.053)	(0.053)	(0.065)	(0.068)	(0.064)	(0.069)	(0.058)	(0.058)
$\chi 2$	4		11		15		4		9		13	
DF	2		5		8		2		5		8	
CFI	0.997		0.992		0.989		0.996		0.992		0.991	
SRMR	0.009		0.039		0.049		0.010		0.030		0.039	
RMSEA [90% CI]	0.049 [0, 0.119]		0.051 [0, 0.095]		0.047 [0, 0.082]		0.052 [0, 0.125]		0.047 [0, 0.094]		0.040 [0, 0.078]	
N	464	372	464	372	464	372	398	372	398	372	398	372

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

Table C.3: Temporal Equivalence 2000 and 2004, Democrats

	2000							2004					
	Equal Form		Equal Factor Loadings		Equal Intercepts			Equal Form		Equal Factor Loadings		Equal Intercepts	
	2000	2016	2000	2016	2000	2016		2004	2016	2004	2016	2004	2016
Past Discrimination	1.000	1.000	1.000	1.000	1.000	1.000	Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—		—	—	—	—	—	—
Deserve Less	0.839	1.058	1.017	1.017	0.967	0.967	Past Discrimination	1.104	0.945	1.003	1.003	0.993	0.993
	(0.130)	(0.080)	(0.068)	(0.068)	(0.059)	(0.059)		(0.110)	(0.071)	(0.060)	(0.060)	(0.056)	(0.056)
Try Hard	0.606	0.763	0.720	0.720	0.750	0.750	Try Hard	0.935	0.722	0.802	0.802	0.835	0.835
	(0.103)	(0.076)	(0.063)	(0.063)	(0.061)	(0.061)		(0.102)	(0.072)	(0.059)	(0.059)	(0.058)	(0.058)
Special Favors	0.458	0.896	0.778	0.778	0.804	0.804	Special Favors	1.056	0.847	0.923	0.923	0.949	0.949
	(0.098)	(0.077	(0.064)	(0.064)	(0.062)	(0.062)		(0.108)	(0.073)	(0.061)	(0.061)	(0.059)	(0.059)
Intercept Past	3.024	2.421	3.024	2.421	3.012	3.012	Intercept Past	2.825	2.421	2.825	2.421	2.870	2.870
Discrimination	(0.085)	(0.084)	(0.082)	(0.085)	(0.075)	(0.075)	Discrimination	(0.064)	(0.068)	(0.063)	(0.069)	(0.057)	(0.057)
Intercept Deserve	3.179	2.737	3.179	2.737	3.255	3.255	Intercept Deserve	3.198	2.737	3.198	2.737	3.214	3.214
Less	(0.076)	(0.081)	(0.076)	(0.081)	(0.071)	(0.071)	Less	(0.075)	(0.084)	(0.074)	(0.084)	(0.068)	(0.068)
Intercept Try Hard	3.106	2.442	3.106	2.442	2.988	2.988	Intercept Try Hard	3.054	2.442	3.054	2.442	2.967	2.967
	(0.083)	(0.085)	(0.083)	(0.085)	(0.071)	(0.071)		(0.075)	(0.085)	(0.073)	(0.088)	(0.066)	(0.066)
Intercept Special	3.618	2.940	3.618	2.940	3.505	3.505	Intercept Special	3.550	2.940	3.549	2.940	3.489	3.489
Favors	(0.081)	(0.088)	(0.086)	(0.085)	(0.073)	(0.073)	Favors	(0.076)	(0.088)	(0.075)	(0.089)	(0.069)	(0.069)
$\chi 2$	0		12		24			1		5		13	
DF	2		5		8			2		5		8	
CFI	1		0.990		0.979			1		1		0.995	
SRMR	0.002		0.053		0.072			0.004		0.030		0.032	
RMSEA [90% CI]	0 [0, 0.058]		0.075 [0.021, 0.129]		0.088 [0.050, 0.130]			0 [0, 0.106]		0.012 [0, 0.083]		0.047 [0, 0.091]	
N	246	275	246	275	246	275		308	275	308	275	308	275

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

C.1.1 Partial Equivalence and Substantive Effects

Table C.5 focuses on temporal equivalence for Democrats. The first panel comparing 2000 and 2016, and so on. I first consider partial equivalence for all models. For the 2000-2016 equal loadings test, modification indices point to *special favors* (MI = 10.01, *p*-value = .006). Its loading is larger in 2016 than 2000 (Expected parameter change $[EPC]_{2000}$ = -.355, EPC_{2016} = .114), an outcome inconsistent with the *measurement* view that measure performance should *decline* rather than improve. Freely estimating this parameter improves model fit (row 3), establishing equal factor loadings. For equal intercepts, modification indices suggest freeing *try hard*'s intercept (MI = 5.65, *p* = .053); the item may underestimate racial resentment in 2016 compared to 2000 (EPC_{2000} = -.092, EPC_{2016} = .116). Row 4 shows that freeing it improves fit, with no reliable change compared to the partial loadings model. The 2000-2016 comparison offers little support for the *measurement* and *expressive* explanations. For 2004, while fit reliably worsens after constraining item intercepts, modification indices do not indicate that freeing any constraints will improve model fit.[6] Equal intercepts is met, supporting the *genuine* view.

[6] The largest MI is 3.03 for *special favors* (*p* = .257).

In panel 3, modification indices suggest freeing *try hard*'s intercept (MI = 11.4, *p*-value = .004). Consistent with the *expressive* explanation it underestimates racial resentment in 2016 compared to 2008 (EPC_{2008} = .127, EPC_{2016} = -.219). Doing so yields a better fitting model (row 4) but one with reliably worse fit on most measures. MIs now point to *special favors* (MI = 12.3, *p*-value = .002). It too underestimates racial resentment in 2016 (EPC_{2008} = .096; EPC_{2016} = -.109). Freeing this parameter improves model fit such that it is indistinguishable from the equal loadings model (row 5), establishing equal intercepts. While the *expressive* explanation receives more support than in other tests, it is limited given partial equivalence.

Comparing 2012 and 2016 suggests potential violation of equal intercepts. Modification indices suggest freeing *try hard* (MI = 5.65, *p* = .053); it may underestimate racial resentment in 2016 compared to 2012 (EPCs .083, -.116). To provide as generous as possible a test of the *expressive* position, I free this item. But as row 4 shows, the model meets partial equivalence.

Table C.4: Temporal Equivalence 2008 and 2012, Democrats

	2008						2012					
	Equal Form		Equal Factor Loadings		Equal Intercepts		Equal Form		Equal Factor Loadings		Equal Intercepts	
	2008	2016	2008	2016	2008	2016	2012	2016	2012	2016	2012	2016
Deserve Less	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	—	—	—	—	—	—	—	—	—	—	—	—
Try Hard	0.668	0.722	0.694	0.694	0.771	0.771	0.804	0.722	0.749	0.749	0.768	0.768
	(0.080)	(0.072)	(0.053)	(0.053)	(0.053)	(0.053)	(0.099)	(0.072)	(0.058)	(0.058)	(0.056)	(0.056)
Special Favors	0.627	0.847	0.740	0.740	0.800	0.800	0.835	0.847	0.840	0.840	0.831	0.831
	(0.070)	(0.073)	(0.052)	(0.052)	(0.051)	(0.051)	(0.097)	(0.073)	(0.058)	(0.058)	(0.055)	(0.055)
Past Discrimination	0.890	0.945	0.918	0.918	0.944	0.944	1.036	0.945	0.972	0.972	0.939	0.939
	(0.089)	(0.071)	(0.055)	(0.055)	(0.050)	(0.050)	(0.111)	(0.071)	(0.060)	(0.060)	(0.054)	(0.054)
Intercept Deserve Less	3.244	2.737	3.244	2.737	3.278	3.278	3.350	2.737	3.350	2.737	3.336	3.336
	(0.056)	(0.081)	(0.056)	(0.081)	(0.054)	(0.054)	(0.064)	(0.081)	(0.064)	(0.080)	(0.062)	(0.062)
Intercept Try Hard	3.250	2.442	3.250	2.442	3.123	3.123	3.074	2.442	3.074	2.442	2.991	2.991
	(0.061)	(0.085)	(0.061)	(0.085)	(0.057)	(0.057)	(0.073)	(0.085)	(0.072)	(0.086)	(0.063)	(0.063)
Intercept Special Favors	3.643	2.940	3.643	2.940	3.563	3.563	3.431	2.940	3.430	2.940	3.432	3.432
	(0.059)	(0.088)	(0.061)	(0.085)	(0.056)	(0.056)	(0.073)	(0.088)	(0.073)	(0.087)	(0.064)	(0.064)
Intercept Past	2.921	2.421	2.921	2.421	2.944	2.944	2.866	2.421	2.866	2.421	2.922	2.922
Discrimination	(0.062)	(0.084)	(0.062)	(0.084)	(0.057)	(0.057)	(0.072)	(0.084)	(0.071)	(0.084)	(0.064)	(0.064)
$\chi 2$	6		12		36		1		2		10	
DF	2		5		8		2		5		8	
CFI	0.997		0.994		0.973		1		1		0.997	
SRMR	0.009		0.035		0.067		0.003		0.014		0.026	
RMSEA [90% CI]	0.071 [0, 0.143]		0.061 [0.013, 0.108]		0.098 [0.067, 0.132]		0 [0, 0.085]		0 [0, 0.048]		0.031 [0, 0.077]	
N	443	275	443	275	443	275	365	275	365	275	365	275

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

Substantively, the *measurement* and *expressive* explanations also appear unable to wholly explain attitude change. For the 2000-2016 test, effect sizes suggest small but possibly meaningful practical effects for inequivalence on the partially equivalent equal intercepts model (*deserve less*: SDI_2 =−.198, UDI_2 = .198. *special favors*: SDI_2 = .161, UDI_2 = .253). Further, respondents in 2000 underreport racial resentment on *deserve less* relative to 2016 which runs against the *expressive* explanation. These effects produce item mean changes of -.29 and .26 units on the 5-point scale and an observed mean difference of .603 (d = .541) largely explained by impact (.635, d = .576) rather than bias (-.032, d = -.035). Inequivalence has somewhat larger consequences for the 2008-2016 comparison (*try hard*: SDI_2 = .321, UDI_2 = .321. *special favors*: SDI_2 = .232, UDI_2 = .232), .543 and .388 point changes in item means. Here the observed mean difference of .637 (d = .581) appears due to bias (.930, d = 1.01) not impact (-.293, d = -.433). While this suggests the *expressive* account may explain observed change, fine performance in other comparisons, and negligible practical effects in the 2000-2016 test, suggest this evidence is limited.

Table C.5: Measurement Equivalence of Racial Resentment by Year, 2016 baseline, Democrats

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	0.239	1.000	0.002	0.000								
Equal Factor Loadings	12.300	0.990	0.053	0.075	12.100	0.008	-0.009	0.013	0.051	0.006	0.075	0.009
Equal Factor Loadings[1]	2.540	1.000	0.020	0.000	2.300	0.290	0.000	0.842	0.017	0.343	0.000	0.798
Equal Intercepts[1]	13.200	0.992	0.040	0.058	10.700	0.012	-0.007	0.018	0.021	0.010	0.058	0.006
Equal Intercepts[1,2]	6.180	1.000	0.031	0.011	3.640	0.164	-0.0002	0.207	0.012	0.073	0.011	0.143
2004												
Equal Form	1.470	1.000	0.004	0.000								
Equal Factor Loadings	5.190	1.000	0.030	0.012	3.720	0.293	-0.0002	0.314	0.025	0.197	0.012	0.212
Equal Intercepts	13.100	0.995	0.032	0.047	7.930	0.046	-0.005	0.035	0.002	0.550	0.035	0.035
2008												
Equal Form	5.600	0.997	0.009	0.071								
Equal Factor Loadings	11.800	0.994	0.035	0.061	6.170	0.098	-0.003	0.107	0.027	0.083	-0.009	0.196
Equal Intercepts	35.700	0.973	0.067	0.098	23.900	0.000	-0.017	0.000	0.032	0.000	0.037	0.018
Equal Intercepts[3]	24.200	0.984	0.059	0.083	12.400	0.004	-0.009	0.004	0.023	0.000	0.021	0.054
Equal Intercepts[3,4]	11.900	0.994	0.036	0.053	0.173	0.664	0.001	0.735	0.0002	0.495	-0.009	0.672
2012												
Equal Form	0.813	1.000	0.003	0.000								
Equal Factor Loadings	2.240	1.000	0.014	0.000	1.420	0.683	0.000	0.884	0.011	0.741	0.000	0.846
Equal Intercepts	10.500	0.997	0.026	0.031	8.220	0.042	-0.002	0.093	0.012	0.081	0.031	0.049
Equal Intercepts[3]	4.810	1.000	0.021	0.000	2.574	0.273	0.000	0.820	0.006	0.198	0.000	0.777

Note: Models use *deserve less* to define the dimension but 2000 where *past discrimination* does. One error covariance estimated between *try hard* and *special favors*. 1: frees *special favors* loading; 2: frees *deserve less* intercept; 3: frees *try hard* intercept; 4 frees *special favors* intercept

C.2 Supplementary Analyses to Study 3

The analyses in Tables C.6-C.9 extend those in Study 3 by replacing 2016 with 2008 or 2012 as the comparison point. It could be the case that the *measurement* or *expressive* explanations hold, but changes occurred prior to 2016 making the tests offered imprecise. Even so, they offer no evidence that using 2016 as the comparison year shifts conclusions. Tables C.6 and C.7 indicate that the racial resentment measure meets equal form, factor loadings, and intercepts for both Democrats and Republicans comparing 2000 and 2004 to 2008. For Democrats, this requires freely estimating *special favors*'s factor loading in the 2004-2008 comparison because it better captures racial resentment in

2004. But this has minimal substantive consequences (SDI_2 = .018, UDI_2 = .145), with the observed mean difference of -.112 scale points explained by impact (-.133, *d* = -.138) and minimal bias (.021, *d* = .025) in opposite directions. For Republicans, the 2000-2008 comparison requires freeing *try hard*'s intercept because of underestimating negative attitudes in 2000 relative to 2008. In contrast to Democrats, this has some substantive consequences (SDI_2 = -0.211, UDI_2 = 0.211), with the observed mean difference of -.149 scale points (*d* = .191) explained by entirely by bias (-.190, *d* = -.265; impact: .041, *d* = .074). These substantive consequences are inconsistent with the *expressive* account.

Table C.6: Measurement Equivalence of Racial Resentment by Year 2008 Baseline, Democrats

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	5.780	0.995	0.010	0.074								
Equal Factor Loadings	10.200	0.993	0.028	0.055	4.400	0.292	-0.002	0.291	0.019	0.325	-0.019	0.306
Equal Intercepts	16	0.989	0.027	0.054	5.790	0.133	-0.004	0.133	-0.002	0.889	-0.001	0.200
2004												
Equal Form	7.010	0.995	0.011	0.082								
Equal Factor Loadings	19.100	0.985	0.050	0.087	12.100	0.021	-0.010	0.017	0.040	0.034	0.005	0.296
Equal Factor Loadings[1]	11.900	0.992	0.031	0.073	4.890	0.124	-0.003	0.104	0.020	0.234	-0.009	0.770
Equal Intercepts[1]	15.200	0.991	0.040	0.056	3.300	0.372	-0.0003	0.282	0.009	0.159	-0.017	0.824

Note: Models use *deserve less* to define the dimension but 2000 where *past discrimination* does. One error covariance estimated between *try hard* and *special favors*. 1: frees *special favors* loading

Table C.7: Measurement Equivalence of Racial Resentment by Year 2008 Baseline, Republicans

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	2.520	0.999	0.008	0.027								
Equal Factor Loadings	5.780	0.999	0.025	0.021	3.260	0.368	-0.001	0.247	0.017	0.588	-0.006	0.868
Equal Intercepts	18.800	0.979	0.046	0.063	13.000	0.009	-0.019	0.006	0.021	0.014	0.041	0.022
Equal Intercepts[1]	9.770	0.995	0.028	0.034	3.990	0.143	-0.004	0.092	0.004	0.399	0.013	0.110
2004												
Equal Form	4.040	0.997	0.009	0.049								
Equal Factor Loadings	10.400	0.992	0.035	0.051	6.350	0.145	-0.005	0.097	0.026	0.288	0.001	0.272
Equal Intercepts	13.900	0.991	0.038	0.042	3.480	0.337	-0.001	0.204	0.004	0.492	-0.009	0.799

Note: Models use *deserve less* to define the dimension but 2000 where *past discrimination* does. One error covariance estimated between *try hard* and *special favors*. 1: frees *try hard* intercept

Table C.8: Measurement Equivalence of Racial Resentment by Year 2012 Baseline, Democrats

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	0.993	1.000	0.005	0.000								
Equal Factor Loadings	7.640	0.996	0.040	0.042	6.650	0.114	-0.004	0.130	0.035	0.073	0.042	0.092
Equal Intercepts	22.900	0.977	0.056	0.078	15.300	0.003	-0.019	0.002	0.016	0.036	0.037	0.033
Equal Intercepts[1]	10.100	0.995	0.044	0.038	2.430	0.295	-0.001	0.197	0.005	0.335	-0.004	0.714
2004												
Equal Form	2.220	1.000	0.006	0.018								
Equal Factor Loadings	4.670	1.000	0.024	0.000	2.450	0.515	0.0003	0.933	0.018	0.520	-0.019	0.927
Equal Intercepts	12.200	0.995	0.019	0.039	7.520	0.060	-0.005	0.046	-0.005	0.991	0.039	0.024
2008												
Equal Form	6.350	0.995	0.010	0.073								
Equal Factor Loadings	9.600	0.995	0.026	0.048	3.250	0.405	-0.0003	0.407	0.016	0.336	-0.026	0.459
Equal Intercepts	23.200	0.984	0.049	0.067	13.60	0.006	-0.011	0.006	0.023	0.002	0.021	0.025
Equal Intercepts[1]	13.100	0.994	0.035	0.046	3.450	0.176	-0.002	0.177	0.009	0.086	-0.002	0.221

Note: Models use *past discrimination* to define the dimension. One error covariance estimated between *try hard* and *special favors*. 1: frees *deserve less* intercept

Similar insights manifest in Tables C.8 and C.9 and the 2012 baseline. Again, the measure meets equal form, factor loadings, and intercepts in all comparisons. For Democrats, this requires freeing *deserve less*'s intercept in the 2000 and 2008 comparisons because it *underestimates racial resentment compared to 2012.* This has small, practically important effects, but they are inconsistent with expectations from alternative explanations (2000: SDI_2 =−.260, UDI_2 = .260; 2008: SDI_2 =−.201, UDI_2 = .201). The observed difference of .046 scale points between 2000 and 2012 comes from this bias (-.324, *d* = -.387) negating impact (.370, *d* = .433) indicating higher on average resentment levels in 2000 than 2012. Same for 2008. The observed difference of .081 scale points between 2008 and 2012 comes from this bias (-.229, *d* = -.268) negating impact (.310, *d* = .349). Substantively, the results are inconsistent with alternative explanations and suggest revisions to observed trends in Democrats average levels of racial resentment. 2012 attitudes may be lower than 2000 and 2008, though not 2004.

Table C.9: Measurement Equivalence of Racial Resentment by Year 2012 Baseline, Republicans

	$\chi 2$	CFI	SRMR	RMSEA	$\Delta\chi 2$	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	2.560	0.999	0.009	0.030								
Equal Factor Loadings	5.470	0.999	0.026	0.017	2.910	0.446	0.0002	0.896	0.017	0.599	-0.013	0.874
Equal Intercepts	17.400	0.978	0.046	0.061	11.900	0.011	-0.021	0.008	0.020	0.016	0.044	0.019
Equal Intercepts[1]	6.950	1.000	0.024	0.000	1.480	0.480	0.001	0.850	-0.002	0.845	-0.017	0.919
2004												
Equal Form	4.080	0.996	0.009	0.052								
Equal Factor Loadings	6.190	0.998	0.022	0.025	2.120	0.600	0.002	0.950	0.012	0.768	-0.027	0.934
Equal Intercepts	19.200	0.981	0.045	0.060	13.000	0.004	-0.016	0.003	0.023	0.003	0.035	0.026
Equal Intercepts[1]	8.780	0.997	0.029	0.026	2.590	0.278	-0.001	0.172	0.007	0.184	0.001	0.164
2008												
Equal Form	1.460	1.000	0.006	0.000								
Equal Factor Loadings	14.700	0.985	0.046	0.067	13.200	0.007	-0.015	0.011	0.040	0.018	0.067	0.002
Equal Factor Loadings[2]	4.620	0.999	0.021	0.019	3.160	0.238	-0.001	0.319	0.015	0.349	0.019	0.047
Equal Intercepts[2]	17.000	0.984	0.038	0.058	2.330	0.372	-0.001	0.339	-0.008	0.885	-0.009	0.610
Equal Intercepts[2,3]	9.880	0.994	0.032	0.039	5.260	0.070	-0.005	0.072	0.011	0.057	0.020	0.055

Note: Models use *past discrimination* to define the dimension. One error covariance estimated between *try hard* and *special favors*. 1: frees *deserve less* intercept; 2: frees *try hard* loading; 3: frees *try hard* intercept

For Republicans, results also require freeing *deserve less*'s intercept for the 2000 and 2004 comparisons because it underestimates racial resentment relative to 2012. Freely estimating *try hard*'s factor loading and intercept is required for 2008. Its factor loading is larger in 2012 and intercept lower in 2012. The inconsistency in item inequivalence does not suggest systematic changes consistent with the *measurement* or *expressive* accounts. Likewise, substantive effects are modest (2000: SDI_2 =−.259, UDI_2 = .259; 2004: SDI_2 =−.214, UDI_2 = .214; 2008: SDI_2 = .161, UDI_2 = .227). For 2000, the observed difference of -.108 scale points (*d* = -.135) is misidentified due to bias (-.237, *d* = -.328) concealing impact in the opposite direction (.129, *d* = .193). This also holds for 2004. The observed difference of -.098 scale points (*d* = -.124) is misidentified due to bias (-.192, *d* = -.264) concealing impact in the opposite direction (.094, *d* = .140). For 2008 a like result holds, but bias and impact change signs. The observed difference of .041 scale points comes from bias (.150, *d* = .199) canceling out impact (-.109, *d* = -.147). While inequivalence manifests, problematic items vary across

years and substantive consequences are mixed. The *measurement* and *expressive* explanations receive weak support.

D Study 4: Expressive Responding and Party Effects

D.1 Main Text Models

Table D.1: Partisanship Equivalence

	Face-to-Face						Web					
	Equal Form		Equal Factor Loadings		Equal Intercepts		Equal Form		Equal Factor Loadings		Equal Intercepts	
	Republicans	Democrats	Republicans	Democrats	Republicans	Democrats	Republicans	Democrats	Republicans	Democrats	Republicans	Democrats
Deserve Less	1	1	1	1	1	1	1	1	1	1	1	1
	—	—	—	—	—	—	—	—	—	—	—	—
Past	0.915	0.945	0.930	0.930	0.945	0.945	1.098	1.053	1.068	1.068	1.070	1.070
Discrimination	(0.132)	(0.071)	(0.062)	(0.062)	(0.046)	(0.046)	(0.071)	(0.045)	(0.038)	(0.038)	(0.029)	(0.029)
Try Hard	0.671	0.722	0.696	0.696	0.826	0.826	0.780	0.843	0.815	0.815	0.911	0.911
	(0.109)	(0.072)	(0.061)	(0.061)	(0.052)	(0.052)	(0.059)	(0.044)	(0.036)	(0.036)	(0.030)	(0.030)
Special Favors	0.517	0.847	0.718	0.718	0.858	0.858	0.697	1.005	0.875	0.875	1.008	1.008
	(0.081)	(0.073)	(0.060)	(0.060)	(0.050)	(0.050)	(0.051)	(0.047)	(0.037)	(0.037)	(0.031)	(0.031)
Intercept	3.895	2.737	3.895	2.737	3.927	3.927	3.937	2.792	3.937	2.792	3.968	3.968
Deserve Less	(0.057)	(0.081)	(0.057)	(0.081)	(0.054)	(0.054)	(0.035)	(0.044)	(0.035)	(0.044)	(0.033)	(0.033)
Intercept Past	3.492	2.421	3.492	2.421	3.537	3.537	3.724	2.505	3.724	2.505	3.763	3.763
Discrimination	(0.068)	(0.084)	(0.067)	(0.084)	(0.060)	(0.060)	(0.041)	(0.048)	(0.040)	(0.048)	(0.037)	(0.037)
Intercept Try	3.627	2.442	3.626	2.442	3.570	3.570	3.585	2.399	3.585	2.399	3.550	3.550
Hard	(0.061)	(0.085)	(0.060)	(0.085)	(0.056)	(0.056)	(0.038)	(0.045)	(0.037)	(0.045)	(0.035)	(0.035)
Intercept	4.167	2.940	4.167	2.940	4.115	4.115	4.145	2.790	4.145	2.790	4.096	4.096
Special Favors	(0.051)	(0.088)	(0.053)	(0.084)	(0.052)	(0.052)	(0.033)	(0.050)	(0.034)	(0.048)	(0.034)	(0.034)
$\chi 2$	3		16		32		1		29		63	
DF	2		5		8		2		5		8	
CFI	0.998		0.986		0.970		1		0.990		0.978	
SRMR	0.008		0.057		0.079		0.001		0.055		0.074	
RMSEA [90% CI]	0.045 [0, 0.128]		0.084 [0.041, 0.130]		0.097 [0.064, 0.133]		0 [0, 0.050]		0.077 [0.051, 0.1043]		0.091 [0.071, 0.112]	
N	372	275	372	275	372	275	899	762	899	762	899	762

Note: Models estimated using maximum likelihood. Parameter estimates with standard errors in parentheses. Error covariance between *try hard* and *special favors* estimated but omitted.

D.2 Partial Equivalence and Substantive Effects

Table D.2: Measurement Equivalence of Racial Resentment by Partisanship

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
Face-to-Face												
Equal Form	3.320	0.998	0.008	0.045								
Equal Factor Loadings	16.300	0.986	0.057	0.084	13.000	0.004	-0.007	0.007	0.049	0.005	0.038	0.057
Equal Factor Loadings[1]	3.470	1.000	0.011	0.000	0.147	0.918	0.001	0.896	0.002	0.925	-0.045	0.986
Equal Intercepts[1]	15.000	0.990	0.042	0.060	11.500	0.008	-0.006	0.020	0.032	0.000	0.060	0.003
Equal Intercepts[1,2]	8.800	0.997	0.026	0.038	5.330	0.077	-0.002	0.103	0.015	0.011	0.038	0.031
Web												
Equal Form	0.671	1.000	0.001	0.000								
Equal Factor Loadings	29.300	0.990	0.055	0.077	28.600	0.000	-0.006	0.001	0.054	0.000	0.077	0.001
Equal Factor Loadings[1]	2.400	1.000	0.013	0.000	1.730	0.434	0.000	0.869	0.012	0.321	0.000	0.828
Equal Intercepts[1]	20.600	0.994	0.036	0.048	18.200	0.002	-0.003	0.003	0.023	0.000	0.048	0.001
Equal Intercepts[1,2]	12.200	0.997	0.025	0.035	9.810	0.005	-0.002	0.015	0.012	0.001	0.035	0.004
Equal Intercepts[1,2,3]	2.400	1.000	0.013	0.000	0.0002	0.988	0.000	0.797	0.00002	0.554	0.000	0.766

Note: Models use *deserve less* to define the dimension. One error covariance estimated between *try hard* and *special favors*.
1: frees *special favors* loading; 2: frees *try hard* intercept; 3: frees *special favors* intercept

Table D.2 shows the measure meets partial equivalence in both samples. In the top panel, modification indices for the equal factor loadings model (row 2) suggest freeing *special favors*'s loading (MI = 13.19, p = .001). But against expectations for the *measurement* explanation, it more strongly relates to racial resentment for Democrats (Expected parameter change [EPC]= .125) than for Republicans (EPC = -.251). The reverse should hold if the *measurement* account explains attitude change. The bottom panel model suggests similar changes. Modification indices point to *special favors* and *past discrimination* as potential sources of worse fit (*Special favors*: MI = 27.21, $p < .001$. *Past discrimination*: MI = 5.67, p = .044). *Special favors*'s relationship with racial resentment is stronger for Democrats than Republicans ($EPC_{Democrats}$ = .125, $EPC_{Republicans}$ = -.199) while *past discrimination* is somewhat less related to racial resentment for Democrats than Republicans ($EPC_{Democrats}$ = -.013, $EPC_{Republicans}$ = .029), inconsistent support for the *measurement* explanation for change. With much larger MI and EPC values, I again free *special favors*'s loading. Freeing these parameters yields models that fit as well as the equal form model, offering little support for the *measurement* explanation.

The test for equal intercepts also suggest limited support for the *expressive* account. In the top panel, evidence suggests freeing *try hard*'s intercept (MI = 6.31, p = .048). It appears to be lower among Democrats (EPC = -.195) than Republicans (EPC = .069), suggesting under-/over-estimation of racial resentment, respectively. Freeing this intercept yields a well-fitting model, but the SRMR and RMSEA still show reliable decreases in model fit (row 5). This may be from marginal issues with *special favors* (MI = 5.40, p = .056). Freeing this intercept may improve model fit because the estimate for Democrats is lower (EPC = -.143; Republicans = .032). But given inconsistent evidence for a decline in model fit, imprecision in whether freeing *special favors*'s intercept is necessary, and great model fit overall, parsimony indicates proceeding with this model as establishing partial equivalence (Bollen 1989).

The bottom panel is similar. Evidence again points to *try hard* (MI = 8.56, p = .010). Its intercept is too high[low] for Democrats[Republicans] (EPCs -.095 and .047), indicating it under(over) estimates racial resentment. Freeing it yields a better fitting model but one that still fits reliably worse than the third row's (row 5). And unlike the face-to-face group, variation is consistent: *special favors* is at issue (MI = 9.99, p = .003). Democrats' intercept is too high relative to Republicans' (EPCs -.111, .023). Freeing it produces a model with fit indistinguishable from the partial equal factor loadings model, establishing partial equal intercepts.

Substantive effects for items contributing to inequivalence, while potentially meaningful, run opposite alternative explanation expectations. For the partially equivalent equal factor loadings models, differences in *special favors*'s factor loading has modest practical effects (Face-to-face: SDI_2 = .307, UDI_2 = .353. Web: SDI_2 = .295, UDI_2 = .323). In the partially equivalent equal intercepts model, *try hard*'s inequivalence has limited practical consequences but *special favors* may have moderate effects (Face-to-face: *special favors*: SDI_2 = .337, UDI_2 = .390, *try hard*: SDI_2 = .172, UDI_2 = .172. Web: *special favors*: SDI_2 = .401, UDI_2 = .407, *try hard*: SDI_2 = .188, UDI_2 = .188). *Special favors*'s inequivalence

produces an over .5 point change in item mean on the 5-point response scale. Non-negligible, but unable to fully explain the partisan attitude gap and, importantly, contrary to expectations from the *measurement* and *expressive* arguments. The observed difference of 1.16 points in the face-to-face group (d = 1.01) is explained mostly by bias (.816, d = .897) than impact (.348, d = .116). And this holds in the web group (observed: 1.23, d = 1.07; bias: .816, d = .860; impact: .409, d = .208). But to emphasize, bias is driven by variation in *special favors*'s factor loading and its worse performance among Republicans, a result inconsistent with any alternative explanation.

D.3 Supplementary Analyses to Study 4

Table D.3: Measurement Invariance of Racial Resentment by Party, 2000-2012

	χ2	CFI	SRMR	RMSEA	$\Delta\chi$2	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
2000												
Equal Form	0.795	1.000	0.003	0.000								
Equal Factor Loadings	3.640	1.000	0.020	0.000	2.850	0.506	0.000	0.865	0.017	0.442	0.000	0.810
Equal Intercepts	18.900	0.987	0.04	0.051	15.200	0.002	-0.011	0.006	0.020	0.001	0.051	0.002
Equal Intercepts[2]	5.130	1.000	0.021	0.000	1.490	0.483	0.000	0.730	0.002	0.507	0.000	0.667
2004												
Equal Form	4.770	0.996	0.010	0.064								
Equal Factor Loadings	28.400	0.970	0.083	0.117	23.600	0.000	-0.020	0.001	0.073	0.000	0.053	0.001
Equal Factor Loadings[1]	8.450	0.994	0.033	0.057	3.680	0.223	-0.002	0.239	0.023	0.185	-0.007	0.234
Equal Intercepts[1]	13.700	0.991	0.048	0.053	5.230	0.170	-0.002	0.193	0.015	0.027	-0.004	0.271
2008												
Equal Form	6.280	0.995	0.009	0.069								
Equal Factor Loadings	16.900	0.986	0.053	0.072	10.600	0.026	-0.006	0.043	0.044	0.007	0.004	0.039
Equal Intercepts	23.400	0.982	0.071	0.065	6.470	0.089	-0.003	0.114	0.017	0.008	-0.007	0.310
2012												
Equal Form	1.530	1.000	0.006	0.000								
Equal Factor Loadings	14.800	0.986	0.050	0.072	13.300	0.007	-0.010	0.018	0.044	0.009	0.072	0.002
Equal Factor Loadings[1]	4.240	1.000	0.020	0.013	2.710	0.319	-0.0002	0.396	0.014	0.385	0.013	0.112
Equal Intercepts[1]	11.900	0.993	0.032	0.043	7.620	0.052	-0.005	0.072	0.012	0.084	0.030	0.034

Note: Models use *past discrimination* to define the dimension but 2008 where *deserve less* does. One error covariance estimated between *try hard* and *special favors*. 1: frees *special favors* loading; 2: frees *special favors* intercept

The analyses in Table D.3 complement those in Study 4 by extending the partisanship equivalence test of the *expressive* explanation for attitude change to face-to-face respondents in other years. The results suggest that the inequivalence identified in the main text is less a function of attitude change coinciding with the *expressive* explanation and more a function of consistent party-based differences. In only one instance does the racial resentment measure meet the full equal form, equal factor loadings, and equal intercepts requirements (2008).[7] In two instances (2004 and 2012) the analyses require freeing *special favors*'s factor loading to establish equal factor loadings and another two comparisons (2000 and 2012) necessitate freeing the item's intercept. But these do not appear to be practically consequential. In all cases effect sizes suggest small but meaningful effects (2000: SDI_2 = .205, UDI_2 = .205. 2004: SDI_2 = .226, UDI_2 = .332. 2012: SDI_2 = .350, UDI_2 = .369) (Gunn, Grimm and Edwards 2019). In 2000, the observed group difference of .461 scale points (d = .494) is explained marginally more by bias (.227, d = .288) than impact (.234, d = .206). This changes slightly in 2004, where the observed difference of .611 points (d = .640) is explained more by impact (.353, d = .346) than bias (.258, d = .294). Only in 2012 do substantive effects resemble those identified in 2016. While the observed group difference of .678 points (d = .688) is substantively the same as 2004, this is explained equally by bias (.311, d = .369) and impact (.367, d = .319). But this is due in particular to *special favors*'s lower factor loading among Republicans, a result inconsistent with expectations from any existing explanation. These results therefore offer additional evidence that observed racial attitude change likely stems more from *genuine* change than other explanations. Methodologically they establish the measure's comparability by party but also suggest areas for improvement given practical consequences from *special favors*.

[7] While all fit measures display reliable decline on the equal factor loadings test, modification indices do not suggest improvement from freely estimating any factor loadings. The largest MI is 5.650 for *past discrimination* but the associated p-value is .071.

E Complementary Evidence

E.1 Muslim American Resentment (MAR)

The main text tests suggest observed changes in Whites' views of Black Americans are more likely *genuine* than due to other explanations. While important, these analyses are limited because they only consider one target group. It could be the case that attitudes about Black Americans are unique. I therefore also consider attitudes about Muslim Americans as captured by Muslim American Resentment (MAR) (Lajevardi 2020). Question wording in Table E.1.

MAR is politically relevant (Collingwood, Lajevardi and Oskooii 2018; Lajevardi and Abrajano 2019) but some of these connections may be shaped by how these attitudes are expressed post-2016 (see Crandall, Miller and White II 2018). I address its comparability and consider the question of identifying genuine versus expressive responding with data Lajevardi and Abrajano (2019) collect as part of the 2016 Comparative Campaign Analysis Project using YouGov's nonrandom respondent pool with completed responses weighted back to national benchmarks.

Table E.1: Question Wording

Item	Question Wording
Integrate	Most Muslim Americans integrate successfully into American culture (R)
Interests	Muslim Americans sometimes do not have the best interests of Americans at heart.
Surveil	Muslims living in the US should be subject to more surveillance than others.
Violent	Muslim Americans, in general, tend to be more violent than other people.
Jihad	Most Muslim Americans reject jihad and violence. (R)
English	Most Muslim Americans lack basic English language skills.
Terrorists	Most Muslim Americans are not terrorists. (R)
Headscarves	Wearing headscarves should be banned in all public places.
Oppose Terrorism	Muslim Americans do a good job of speaking out against Islamic terrorism. (R)

Note: (R) denotes reverse coding. Responses recorded on 6-point strongly disagree—strongly agree scales.

I compare responses to the MAR measure by party to test the *expressive* explanation. Like the racial resentment analyses, it could be that party-specific norms shape response patterns. However, these data can only assess weak and strong partisans. Replication data from Lajevardi and Abrajano (2019) lack the full branched survey item, so independent leaners are excluded.

Table E.2's results show that the measure meets equal form (row 1) but equal factor loadings fails (row 2). Modification indices suggest freeing *English* (MI = 15.1, $p < .001$). Its loading is larger for Democrats than Republicans (EPCs .081 and -.189). Freeing this yields a model whose fit does not differ from the equal form model (row 3). MAR meets equal factor loadings. The equal intercepts test also fails. *Violent* and *oppose terrorism* contribute to error (*violent*: MI = 8.96, $p = .026$. *oppose terrorism*: MI = 26.6, $p < .001$.). Comparing EPCs, freeing *oppose terrorism*'s intercept appears to contribute the most to model fit improvement (*violent*: EPC Democrats = .073, Republicans = -.204; *oppose terrorism*: Democrats = -.292, Republicans = .427). Freeing *oppose terrorism* does yield an improvement in model fit (row 5), but it is still reliably worse than the partial equal factor loadings model in row 3. Now the item at issue is *integrate* (MI = 21.2, $p < .001$). Unlike *oppose terrorism*, it appears to underestimate Democrats' attitudes (EPC_{Dem} -.128; EPC_{Rep} = .273). Freeing this item's intercept alongside *oppose terrorism*'s establishes equal intercepts (row 6).

Table E.2: Measurement Equivalence of Muslim American Resentment by Party

	$\chi 2$	CFI	SRMR	RMSEA	$\Delta\chi 2$	p-value	ΔCFI	p-value	ΔSRMR	p-value	ΔRMSEA	p-value
Equal Form	124	0.967	0.033	0.088								
Equal Factor Loadings	148	0.960	0.062	0.088	24.3	0.012	-0.005	0.017	0.029	0.005	0.001	0.009
Equal Factor Loadings[1]	133	0.966	0.047	0.082	9.07	0.353	-0.001	0.364	0.014	0.197	-0.005	0.287
Equal Intercepts[1]	194	0.944	0.083	0.098	61.3	0.000	-0.016	0.000	0.037	0.000	0.016	0.000
Equal Intercepts[1,2]	166	0.955	0.062	0.088	33.5	0.000	-0.008	0.000	0.016	0.000	0.006	0.000
Equal Intercepts[1,2,3]	145	0.963	0.053	0.080	12	0.064	-0.002	0.076	0.006	0.006	-0.002	0.049

Note: Models use *surveil* to define the dimensions. Error covariances estimated between reverse-worded items. 1: frees *english* loading; 2: frees *oppose terrorism* intercept; 3: frees *integrate* intercept

The results do not support the *expressive* account where partisans' use of the MAR measure systematically varies. Three items do exhibit inequivalence, but nothing appears to systematically connect item content to suggest substantive reasons for inequivalence. *English* and *integrate* appear similar in that they relate to acculturation concerns, but a related item– *headscarves*–performs well. Likewise, *oppose terrorism*'s issues do not manifest on similar items (*jihad*, *terrorists*). While the potential issue with *violent* in the equal intercepts test is suggestive, it is not conclusive. Party differences in attitudes about Muslims appear *genuine*.

E.2 Motivations to Control Prejudice

I also complement the main text results by tracking movement in Whites' motivation to control prejudice over time. If these motivations exhibit changes similar to self-reported racial attitudes, then this suggests that some of these observed changes may be due to changing response pressures suggested by the *socially desirable* explanation. Further, if this varies by political orientation, then this speaks to the *expressive* argument. Supportive evidence comes from external motivations increasing over time with internal not changing or decreasing. Self presentation, not personal commitment, appears more at play. Inconsistent evidence would come from internal motivations increasing and no alteration for external. Changes in personal commitment, not evaluations of others, likely better correspond with shifting attitudes.

Unfortunately publicly available representative data collections like the ANES or General Social Survey do not contain such measures. Therefore to understand this I turn to data from Harvard's Project Implicit collected 2015-2018. Project Implicit describes itself as a "'virtual laboratory' for collecting data on the Internet." Importantly, these data are not a random sample. Participants opt-in to participating in a study, with most studies involving completing some form of the Implicit Association Test (Greenwald, McGhee and Schwartz 1998). Alongside these tests, participants report various demographics and answer a variety of self-report measures. Two of these batteries are Plant and Devine's (1998) internal and external motivations to respond without prejudice (MCP, featured in Table E.3). I use these to capture motivations.

Table E.3: Motivation to Respond Without Prejudice Question Wording

External Motivation	Internal Motivation
Because of today's PC (politically correct) standards, I try to appear nonprejudiced.	I attempt to act in nonprejudiced ways because it is personally important to me.
I try to hide any negative prejudicial thoughts in order to avoid negative reactions from others.	According to my personal values, using stereotypes is OK. (R)
If I acted prejudiced, I would be concerned that others would be angry with me.	I am personally motivated by my beliefs to be nonprejudiced
I attempt to appear nonprejudiced in order to avoid disapproval from others.	Because of my personal values, I believe that using stereotypes is wrong.
I try to act nonprejudiced because of pressure from others.	Being nonprejudiced is important to my self-concept.

Note: (R) denotes reverse coding. Responses recorded on 11-point very strongly disagree—very strongly agree scales.

While a unique, non-representative sample, I use these data to gain some insight into whether these motivations exhibit trends paralleling other self-reports. To do this I use data from all non-Hispanic White US completes for 2015-2018, years in which the MCP measures are available. Further, I use 2015 as a baseline year for observables and weight 2016-2018 to the 2015 distribution of sex, education, age, region, reason for visiting the website, and ideological self-identification through rake weights. This attempts to as best as possible hold constant the distribution of types of individuals completing an IAT based on recorded observables. While not intended to reflect the US population, this approach can speak to whether a consistent sample of similar individuals reveals changes over time, an approach others have used for similar analyses of opt-in, cross-sectional samples (Clinton, Engelhardt and Trussler 2019).

I then calculate yearly averages for internal and external MCP as well as affective ratings of Black and White people captured on 11-point scales. Further, because Project Implicit does not measure partisanship I also break down trends by ideological self-identification, an imperfect but useful substitute given the increasing alignment between the two (Levendusky 2009).

These data indicate considerable stability in external MCP with some suggestion that internal MCP strengthened. From 2015-2018 internal motivations average .75, .77, .78, and .79 on a 0-1 scale. This varies by ideology, with moderates and conservatives reporting stronger motivations after 2015 (Moderate: .71, .72, .74, and .76; Conservative: .67, .71, .72, and .73). Liberals exhibit much more stability (.81, .80, .82, and .82). External motivations evince greater aggregate (.52, .51, .50, .51) and group-level stability (Liberal: .52, .50, .49, and .50; Moderate: .52, .49, .49, and .52; Conservative: .50, .56, .54, and .52). This suggests that observed changes in explicit attitudes are not necessarily connected to shifting external motivations to respond without prejudice, though unfortunately I cannot extend this trend before 2015 to strengthen this possibility. Further, the trends in internal motivations for moderates and conservatives are consistent with the *genuine* position where increased personal motivation to not respond with prejudice matters. These trends, further, sync with trends in stereotyping Hopkins and Washington (2020) report. Some suggestion of increased external motivations among conservatives does imply altered responding due to social desirability. But if this

exists, then the consequences would suggest sharper changes in other self-reported racial attitudes among this group than data reveal.

Complementing these patterns, I find more favorable trends in group affect, though changes are modest. To account for individual differences in the use of the feeling thermometer items I subtract Whites' ratings of Blacks from their ratings of Whites and set this to run from 0-1, with higher values denoting greater relative preference for Whites. Between 2015 and 2018 this measure reveals a slight decrease: .53, .52, .51 .51. This decline, further, occurs across all ideological groups, though to varied degrees (Liberal: .51, .51, .50, and .50; Moderate: .53, .52, .52, and .52; Conservative: .55, .54, .54, and .53). Whites report somewhat less relative preference for White over Black Americans.

Fortunately, I can also extend this affect measure prior to 2015. These trends are consistent with patterns reported in nationally representative samples: stability through around 2013 then positive shifts. From 2008-2014, these are .54, .55, .55, .55, .54, and .53. This stability is also reflected by political orientation (Liberal: .53 through 2013 then .52 on 2014; Moderate: .54 all; Conservative: .57 through 2013 then .56) Though not as dramatic as changes for other operationalizations reported elsewhere (e.g., Engelhardt 2019), the trends are consistent.

I also investigated whether the correlations between affective ratings and internal and external motivations change over time. If so, then this can shed additional light on the observed trends. To do this I regress the differenced feeling thermometer on each of these dimensions as well as demographics, ideological self-identification, region, and the reason respondents report for navigating to Project Implicit, all the variables used to generate the weights (Berinsky 2009). I run these models separately by year and also stack these four data sets and interact motivations with a year indicator to see if any changes in correlations are reliable.

The results, reported in Table E.4, point to an association between motivations and affect ratings. But they also offer no clear evidence for changes over time. Correlations change relative to 2015, but they are indistinguishable from 0. While again a limited timeframe, these associations suggest changes in motivations do not necessarily underpin self-reports. While it cannot speak to *increased* reliance on internal motivations, they also do not suggest a renewed influence for external. More

generally, internal motivations are consistently more influential than external, a result suggesting self-reports are somewhat more related to internalized concerns rather than perceived external pressures.

I also break this out by ideological self-identification. I look within-year to see if motivations vary in their influence by political orientation in ways suggesting other concerns may intrude on the reporting of racial attitudes. Table E.5 reports these results. They reveal no emergent gaps between liberals and moderates and conservatives in the association between either dimension and affect over time. Internal motivations have stronger associations with affect ratings for conservatives than liberals in 3 of 4 years, but the size of this difference in consistent. Further, changes in internal motivations would suggest attitudes more strongly related to a personal desire to not be prejudiced, an outcome consistent with the *genuine* argument.

Although drawing on a unique sample, and beginning only in 2015, the analyses presented here are consistent with the insights from the main text analyses. Internal, not external, motivations appear to change over time, with moderates and conservatives more personally motivated to not express prejudice. Further, the associations between these motivations and affective evaluations of Black and White Americans are not patterned in ways supporting the *socially desirable* or *expressive* theories of change.

Table E.4: Motivation to Control Prejudice and Pro-White Affect

	2015	2016	2017	2018	All
External Motivations	.040**	.073***	.049***	.065***	.046**
	(.017)	(.015)	(.007)	(.005)	(.019)
Internal Motivations	-.079***	-.092***	-.071***	-.085***	-.063***
	(.018)	(.016)	(.008)	(.006)	(.019)
Moderate	-.002	.014*	.006	.005	.006**
	(.008)	(.008)	(.004)	(.003)	(.002)
Conservative	.019**	.010	.022***	.023***	.021***
	(.009)	(.007)	(.003)	(.003)	(.002)
Male	-.018***	-.002	.004	.003	.002
	(.007)	(.006)	(.003)	(.002)	(.002)
Some college	-.008	.013	-.002	.002	.001
	(.014)	(.009)	(.004)	(.004)	(.003)
College degree	-.007	.011	.001	-.003	-.001
	(.015)	(.009)	(.004)	(.004)	(.003)
30-44	-.013	-.009	-.007*	-.004	-.006***
	(.009)	(.008)	(.004)	(.003)	(.002)
45+	-.002	.005	-.001	-.0004	-.0001
	(.010)	(.008)	(.004)	(.003)	(.002)
Northeast	-.007	-.016**	.001	-.005*	-.004*
	(.009)	(.008)	(.004)	(.003)	(.002)
South	-.013	-.006	-.002	-.001	-.002
	(.009)	(.007)	(.004)	(.003)	(.002)
West	-.015*	-.015*	-.006*	-.006**	-.007***
	(.009)	(.008)	(.004)	(.003)	(.002)
News mention	-.008	-.009	-.007	-.001	-.006
	(.009)	(.009)	(.006)	(.006)	(.004)
Other reason	.007	-.014	.007	-.004	-.001
	(.012)	(.012)	(.008)	(.006)	(.004)
Peer mention	.007	-.015*	-.001	-.009*	-.005
	(.009)	(.009)	(.005)	(.005)	(.003)
External*2016					.022
					(.024)
External*2017					.004
					(.020)
External*2018					.020
					(.020)
Internal*2016					-.025
					(.025)
Internal*2017					-.008
					(.021)
Internal*2018					-.024
					(.020)
2016					.004
					(.023)

2017					-.003
					(.019)
2018					.002
					(.019)
Constant	.580***	.545***	.538***	.544***	.543***
	(.023)	(.017)	(.008)	(.007)	(.018)
Observations	490	669	2,735	4,416	8,310
R^2	.106	.123	.088	.107	.099
Residual Std. Error	.068	.074	.077	.084	.080

Note: *p<.01; **p < .05; ***p<0.001 Non-Hispanic Whites. Variables scaled 0-1 or entered as indicators. 2016, 2017, and 2018 weighted to match distribution of observables in 2015. Omitted categories are liberal, HS or less, 18-29, Midwest, and assignment for work or school.

Table E.5: Motivation to Control Prejudice and Pro-White Affect by Ideology

	2015	2016	2017	2018
External Motivations	.028	.066***	.051***	.066***
	(.022)	(.017)	(.008)	(.007)
--*Moderate	.033	-.011	-.004	-.018
	(.039)	(.039)	(.020)	(.016)
--*Conservative	.007	.062	-.003	.006
	(.043)	(.038)	(.015)	(.013)
Internal Motivations	-.064**	-.089***	-.050***	-.066***
	(.027)	(.021)	(.010)	(.009)
--*Moderate	.021	.033	-.022	.020
	(.041)	(.045)	(.022)	(.018)
--*Conservative	-.089**	-.039	-.063***	-.078***
	(.044)	(.039)	(.018)	(.015)
Moderate	-.033	-.004	.026	.001
	(.037)	(.038)	(.019)	(.016)
Conservative	.077*	.004	.071***	.078***
	(.040)	(.033)	(.016)	(.013)
Male	-.018***	-.001	.004	.002
	(.007)	(.006)	(.003)	(.002)
Some college	-.009	.013	-.002	.002
	(.014)	(.009)	(.004)	(.004)
College degree	-.008	.011	.001	-.002
	(.015)	(.009)	(.004)	(.004)
30-44	-.014	-.007	-.007*	-.003
	(.009)	(.008)	(.004)	(.003)
45+	-.002	.005	-.001	.0002
	(.010)	(.008)	(.004)	(.003)
Northeast	-.008	-.016**	.001	-.006*
	(.009)	(.008)	(.004)	(.003)
South	-.016*	-.006	-.002	-.001
	(.009)	(.007)	(.004)	(.003)
West	-.015	-.015*	-.007*	-.006**
	(.009)	(.008)	(.004)	(.003)
News mention	-.007	-.009	-.007	-.001
	(.009)	(.009)	(.006)	(.006)
Other reason	.006	-.015	.007	-.005
	(.012)	(.012)	(.008)	(.006)
Peer mention	.006	-.015*	-.001	-.010**
	(.009)	(.009)	(.005)	(.005)
Constant	.576***	.546***	.520***	.526***
	(.029)	(.020)	(.010)	(.009)
Observations	490	669	2,735	4,416
R^2	.119	.129	.093	.114
Residual Std. Error	.068	.074	.077	.084

Note: *p<.01; **p < .05; ***p<0.001 Non-Hispanic Whites. Variables scaled 0-1 or entered as indicators. 2016, 2017, and 2018 weighted to match distribution of observables in 2015. Omitted categories are liberal, HS or less, 18-29, Midwest, and assignment for work or school.

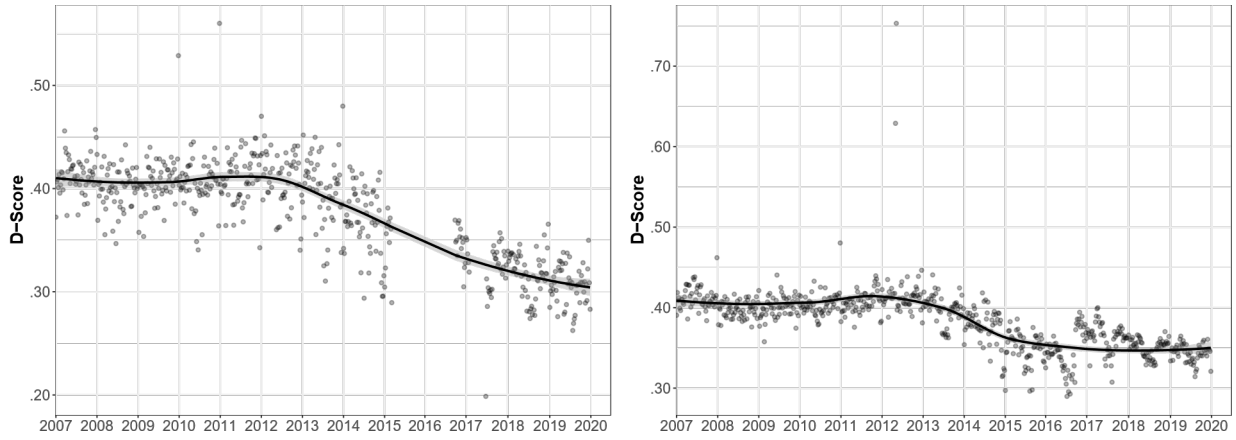
E.3 Evidence from the IAT

I also take advantage of Project Implicit's data on the Black-White IAT to understand whether a measure of implicit attitudes reveals trends like survey self-reports. As described in the main text, if implicit attitudes change then I view this as evidence corroborating my conclusions. As orientations distinct from racial resentment and other explicit attitudes, implicit attitudes are less susceptible to, though not wholly isolated from, the responding motivations featured in the *socially desirable* and *expressive* explanations (Greenwald and Lai 2020). Nor can *measurement* explain trends because the IAT captures affective associations through a specific task, removing changes in measure interpretation as an explanation.

I use data from 2007-2019 and apply the same weighting procedure used in the prior section to describe trends in motivation to control prejudice. Here I weight to week instead of year for increased granularity. I select a week at random in 2007 for this baseline. I plot these averages in Figure E.1, with panel (a) using the weighted sample and (b) the raw data. I add a smoothed loess trend to highlight changes.

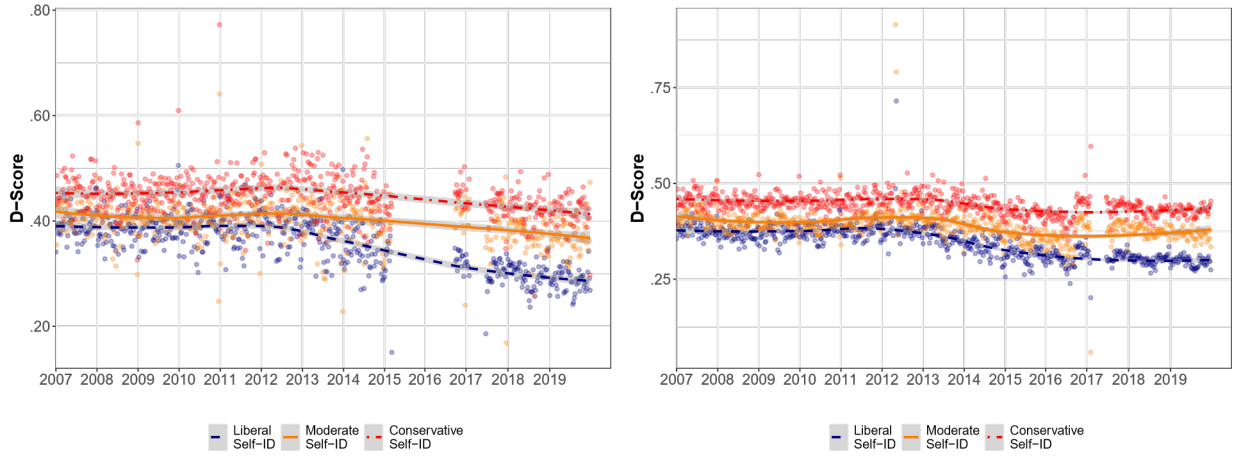
Between the week starting January 1, 2007, and the one ending December 31, 2019, Whites' IAT D-scores (scored -1, 1) decline from .41 to .30 according to estimates from the loess trend using the weighted samples (Figure E.1a). Moreover, this decline in pro-White bias begins after 2012 like the self-reports. Nor does this insight appear due to the weighting procedure used to construct consistent samples. Average D-scores change from .41 to .35 in the raw data.

While magnitudes vary by ideological self-ID, trends do not. Figure E.2 reports this. According to the descriptives in panel (a), liberals shift from .39 to .29, moderates .42 to .37, and conservatives from .45 to .41. Raw score averages reveal similar trends. In panel (b), liberals decrease from .38 to .30, moderates from .41 to .38, and conservatives from .46 to .44.



(a) Weighted (b) Unweighted

Figure E.1: Weekly average IAT D-scores with smoothed trend line. Weeks with missing data are due to missingness on one or more weighting variables. Higher D-Scores indicate greater pro-White bias.



(a) Weighted (b) Unweighted

Figure E.2: Weekly average IAT D-scores with smoothed trend line. Weeks with missing data are due to missingness on one or more weighting variables.

These results complement the conclusions from the main text analyses. While descriptive, the trends parallel those found in self-reports. That these trends manifest on an operationalization of racial attitude consistently captured by response latencies rather than survey questions suggests the *measurement* explanation is unlikely at play. Further, while potentially contributing to responses via introspection (Greenwald and Lai 2020), the results suggest the *socially desirable* and *expressive* positions unlikely explain observed trends. These motivations likely influence the IAT task less. Importantly, while supportive descriptives, these trends require a more exhaustive investigation to

explain and understand. I point them out simply as patterns consistent with the findings in the main text.

E.4 Non-linear Confirmatory Factor Analysis

The multi-group confirmatory factor analysis (MGCFA) approach I use as my focal analytical strategy makes two assumptions. First, is assumes a linear mapping between latent racial attitude and observed item responses. Second, is says that this mapping varies systematically across types of individuals where certain response motivations may differ to explore evidence for a given attitude change explanation. But, importantly, response motivations may vary not just by type of individual but also by trait level. The most prejudiced, for instance, have more to gain from disguising their true attitudes if social desirability pressures have changed and made prejudiced views less socially acceptable. There's thus a non-linear relationship in response outcomes based on someone's attitude level, something the model I use simplifies away from.

To understand if attitude level plays such a role I complement the main text analyses by running a series of non-linear confirmatory factor analyses (NLCA, McDonald 1967). While the model remains linear in parameters, it now includes latent attitude and its square. The rate of change for observed item response for a given increase in latent racial resentment can vary at an increasing or decreasing rate, more flexibly allowing for variation in response pressures by trait level. A limitation of this approach is that to my knowledge it lacks established and validated procedures for comparing groups that I use for MGCFA. Given this, I view these results as exploratory, striving for clarifying the focal analyses.

For these analyses I focus on the comparisons made in studies 1 and 2, the tests of the *socially desirable* explanation. That trait level likely intersects most with this explanation and the most prejudiced being most motivated to edit their responses makes these a most likely place to see if trait level variation matters. I estimate NLCFA models separately for each interview mode, an analysis analogous to the equal form step of the MGCFA approach.[8] I do so to see if any variation in trait level

[8] Analyses conducted with Mplus version 8.3

effects systematically differ by interview context. If so, it suggests that the conclusions from the main text analyses are more limited because they do not reveal the consequences of these motivations. If not, then while variation in trait level may exist, these effects occur to similar degrees by group. Consequently, the MGCFA's focus on types of individuals, rather than trait level, provides a good accounting of varied response pressures.

Table E.6 provides the parameter estimates for the NLCFA models relevant to Study 1. The left column contains the results for the face-to-face respondents. The right column the web respondents. I focus first on the estimates for the face-to-face group. The first panel reports the conventional factor loadings reporting how much change in observed responses on an item comes from a shift in latent racial resentment. The second panel reports the loadings on the quadratic terms. The signs indicate that the link between latent racial resentment and observed item responses weakens as racial resentment increases; observed racial resentment increases but at a decreasing rate. This suggests that there's some variation by trait level.

But comparing across models, the parameter estimates are remarkably similar. Factor loadings and intercepts diverge minimally. Nor does the quadratic effect of latent racial resentment on observed scores differ much. These results suggest trait-level variation occurs no matter the context. The conclusions drawn from the MGCFA approach appear to be a fair characterization of limited response pressures due to social desirability and trait-level variation in effects.

The results reported in Table E.7 offer like insight for Study 2. Comparing columns 1 and 2 and 3 and 4 sees little difference in factor loadings or intercepts across mode within Democrats or Republicans, respectively. Nor do the quadratic effects vary much with Republicans in particular displaying remarkable consistency. Two of these parameters do vary for Democrats. There appear to be no quadratic effects on *try hard* and *special favors* within the web sample but reliable relationships in the face-to-face sample. This does suggest some contribution for scores on latent racial resentment. But at the same time, the substantive consequence is that changes in racial resentment are *increasingly* consequential for observed response among the face-to-face group, an outcome opposite expectations if more resentful individuals constrain their responses more. Variation in

observed responses should be decreasingly associated with an increase in latent racial resentment. While influential, there do not seem to be systematic effects by trait level that change the insights from the MGCFA approach.

Table E.6: Study 1: Mode Comparison

	Face-to-Face	Web
Factor Loadings: Latent Racial Resentment		
Deserve Less	1.000	1.000
	—	—
Try Hard	1.114	1.008
	(.088)	(.098)
Special Favors	1.210	1.161
	(.099)	(.131)
Past Discrimination	1.105	1.216
	(.077)	(.084)
Factor Loadings: Latent Racial Resentment^2		
Deserve Less	-.175	-.162
	(.026)	(.022)
Try Hard	-.228	-.144
	(.036)	(.047)
Special Favors	-.602	-.402
	(.035)	(.112)
Past Discrimination	-.112	-.169
	(.033)	(.048)
Intercepts		
Deserve Less	3.540	3.563
	(.067)	(.043)
Try Hard	3.309	3.181
	(.066)	(.057)
Special Favors	4.157	3.913
	(.062)	(.089)
Past Discrimination	3.111	3.317
	(.075)	(.064)
N	716	1912
AIC	8714	22036
BIC	8792	22130

Note: Models estimated using robust maximum likelihood. Parameter estimates with standard errors in parentheses.

Table E.7: Study 2: Mode Comparison within Party

	Democrats		Republicans	
	Face-to-Face	Web	Face-to-Face	Web
Factor Loadings: Latent Racial Resentment				
Deserve Less	1.000	1.000	1.000	1.000
	—	—	—	—
Try Hard	.894	.968	.880	.744
	(.100)	(.063)	(.083)	(.038)
Special Favors	1.030	1.129	1.002	1.029
	(.101)	(.063)	(.038)	(.001)
Past Discrimination	1.007	1.085	.700	.693
	(.062)	(.043)	(.085)	(.043)
Factor Loadings: Latent Racial Resentment^2				
Deserve Less	.176	.113	-.182	-.230
	(.035)	(.024)	(.022)	(.015)
Try Hard	.128	.162	-.183	-.190
	(.053)	(.034)	(.040)	(.025)
Special Favors	.016	.081	-.542	-.585
	(.048)	(.039)	(.015)	(.000)
Past Discrimination	.269	.215	-.021	-.135
	(.047)	(.032)	(.043)	(.033)
Intercepts	2.522	2.683	3.931	3.985
Deserve Less	(.106)	(.061)	(.065)	(.036)
Try Hard	2.285	2.242	3.678	3.640
	(.125)	(.066)	(.065)	(.038)
Special Favors	2.920	2.713	4.516	4.553
	(.137)	(.081)	(.046)	(.001)
Past Discrimination	2.092	2.296	3.424	3.733
	(.113)	(.070)	(.079)	(.047)
N	275	762	372	899
AIC	3347	8700	4301	7794
BIC	3405	8774	4360	7866

Note: Models estimated using robust maximum likelihood. Parameter estimates with standard errors in parentheses.

E.5 Hypothetical Maximum Effects

I also investigated the greatest extent to which each explanation may explain observed patterns. To do this I calculate Gunn, Grimm and Edwards's (2019) SDI_2 effect size metric for the equal form and equal factor loadings models estimated in each study. These offer the greatest possible effect for the *measurement* and *socially desirable* or *expressive* explanations. Further, because the equal form model frees loadings and intercepts, while the equal loadings just frees intercepts, divergence in effects across the two speaks to contributions from the *measurement* explanation. Consistency says that divergent intercepts matter, the contribution from either the *socially desirable* or *expressive* views.

I report the results of this exercise in Table E.8. I focus first on panel 1 and the maximum contribution the *measurement* and *socially desirable* explanations might have in Study 1. The results suggest scant contribution to observed scores. The numeric entries report the average absolute value for SDI_2 effect sizes of the four racial resentment items. The scale aligns with Cohen's *d*, so I use .20 as a benchmark for small but meaningful effects in this exploratory analysis.[9] With an average of .058, there is likely limited effect on observed responses. Further, that this average holds for both maximal investigations says that the *socially desirable* explanation, not *measurement*, accounts for this negligible divergence.

The second panel offers similar insight. For both Democrats and Republicans, consistency in average effect sizes between tests indicates no meaningful contribution from *measurement* to observed responses. Further, with average effects sizes of .062 and .071 the *socially desirable* theory of change has negligible substantive effects.

The third panel offers continued evidence for *measurement* likely not mattering, but with larger substantive effects from *expressive*. The comparisons of Republicans' responses over time all suggest negligible average effects, with the largest average of .100 in the equal loadings and intercepts tests

[9] This provides a more restrictive comparison than the .4, .6, .8 of small, medium, and large effects proposed for some equivalence measures (Nye et al. 2019). Gunn, Grimm and Edwards (2019) do not report similar benchmarks for SDI_2, so I use these benchmarks as suggestive for narrative rather than definitive. Comparisons across models are most instructive.

comparing 2008 and 2016. Importantly, no *measurement* contribution appears for any tests, and with average effect sizes of .070, .073, and .65 for 2000, 2004, and 2012, the *expressive* explanation appears to contribute little.

The effects for Democrats rule out meaningful *measurement* contributions with consistent results between tests. But the effect sizes range from .382 to .465, suggesting likely at most moderate practical effects from *expressive*. If the *expressive* position received empirical support, then much of the difference in Democrats' responses to the racial resentment measure in 2016 could be attributed to this relative to the *genuine* explanation.

The results in panel 4 again rule out much contribution from *measurement*. Average effect sizes do not change between tests. Importantly, though, they indicate that the *expressive* explanation could explain much of the difference in racial resentment between Democrats and Republicans if supported statistically. At .840 and .915, average effect sizes are quite large.

These results therefore suggest that the *expressive* position has the greatest potential to explain observed patterns. The *measurement* explanation appears to have scant influence, even at best. There is some potential contribution for *socially desirable*, but nothing removing the contribution of *genuine* change.

Table E.8: Average Inequivalence Effect Sizes by Test

Study		Test	Mean SDI_2s
1		Equal Loadings	0.058
		Equal Intercepts	0.058
2	Democrats	Equal Loadings	0.062
		Equal Intercepts	0.062
	Republicans	Equal Loadings	0.071
		Equal Intercepts	0.071
3	Republicans 2000-2016	Equal Loadings	0.070
		Equal Intercepts	0.070
	2004-2016	Equal Loadings	0.073
		Equal Intercepts	0.073
	2008-2016	Equal Loadings	0.100
		Equal Intercepts	0.100
	2012-2016	Equal Loadings	0.065
		Equal Intercepts	0.065
	Democrats 2000-2016	Equal Loadings	0.432
		Equal Intercepts	0.432
	2004-2016	Equal Loadings	0.382
		Equal Intercepts	0.382
	2008-2016	Equal Loadings	0.465
		Equal Intercepts	0.465
	2012-2016	Equal Loadings	0.392
		Equal Intercepts	0.392
4	Face-to-Face	Equal Loadings	0.840
		Equal Intercepts	0.840
	Web	Equal Loadings	0.915
		Equal Intercepts	0.915

Note: Effect sizes from assuming full inequivalence for loadings and intercepts tests.

References

Berinsky, Adam J. 2009. *In Time of War*. Chicago: University of Chicago Press.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Brown, Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research*. 2 ed. New York: Guilford Press.

Byrne, Barbara M, Richard J Shavelson and Bengt Muthen. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105(3):456–466.

Clinton, Joshua D, Andrew M Engelhardt and Marc J Trussler. 2019. "Knockout Blows or the Status Quo? Momentum in the 2016 Primaries." *The Journal of Politics* 81(3):997–1013.

Collingwood, Loren, Nazita Lajevardi and Kassra A R Oskooii. 2018. "A Change of Heart? Why Individual-Level Public Opinion Shifted Against Trump's "Muslim Ban"." *Political Behavior* 40(4):1035–1072.

Crandall, Christian S, Jason M Miller and Mark H White II. 2018. "Changing Norms Following the 2016 U.S. Presidential Election." *Social Psychological and Personality Science* 9(2):186– 192.

Engelhardt, Andrew M. 2019. "Trumped by Race: Explanations for Race's Influence on Whites' Votes in 2016." *Quarterly Journal of Political Science* 14(3):313–328.

Greenwald, Anthony G and Calvin K Lai. 2020. "Implicit Social Cognition." *Annual review of psychology* 71(1):419–445.

Greenwald, Anthony G, Debbie E McGhee and Jordan LK Schwartz. 1998. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of Personality and Social Psychology* 74(6):1464–1480.

Gunn, Heather J, Kevin J Grimm and Michael C Edwards. 2019. "Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes." *Structural Equation Modeling: A Multidisciplinary Journal* .

Hopkins, Daniel J and Samantha Washington. 2020. "The Rise of Trump, the Fall of Prejudice? Tracking White Americans' Racial Attitudes 2008-2018 via a Panel Survey." *Public Opinion Quarterly* 84(1):119–140.

Jorgensen, Terrence D, Benjamin A Kite, Po-Yi Chen and Stephen D Short. 2018. "Permutation Randomization Methods for Testing Measurement Equivalence and Detecting Differential Item Functioning in Multiple-Group Confirmatory Factor Analysis." *Psychological Methods* 23(4):708–728.

Lajevardi, Nazita. 2020. *Outsiders at Home*. New York: Cambridge University Press.

Lajevardi, Nazita and Marisa Abrajano. 2019. “How Negative Sentiment toward Muslim Americans Predicts Support for Trump in the 2016 Presidential Election.” *The Journal of Politics* 81(1):296–302.

Levendusky, Matthew S. 2009. *The Partisan Sort*. Chicago: University of Chicago Press.

McDonald, Roderick P. 1967. *Nonlinear Factor Analysis (Psychometric Monograph No. 15).* Richmnod: Psychometric Corporation.

Nye, Christopher D and Fritz Drasgow. 2011. “Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups.” *Journal of Applied Psychology* 96(5):966–980.

Nye, Christopher D, Jacob Bradburn, Jeffrey Olenick, Christopher Bialko and Fritz Drasgow. 2019. “How Big Are My Effects? Examining the Magnitude of Effect Sizes in Studies of Measurement Equivalence.” *Organizational Research Methods* 22(3):678–709.

Plant, E Ashby and Patricia G Devine. 1998. “Internal and external motivation to respond without prejudice.” *Journal of personality and social psychology* 75(3):811–832.