ANES Scales Often Don't Measure What You Think They Measure *

MATTHEW T. PIETRYKA
mpietryka@fsu.edu
Associate Professor
Department of Political Science
Florida State University

RANDALL C. MACINTOSH
rmacintosh@csus.edu
Emeritus Professor
Department of Sociology
California State University, Sacramento

January 18, 2021

This paper is forthcoming at the *Journal of Politics*, expected publication date around April, 2022

*This project began as part of the 2016 Election Research Preacceptance Competition (ERPC). For thoughtful suggestions, we thank Doug Ahler, Charles Barrilleaux, Quintin Beazer, Inken von Borzyskowski, Rob Carroll, Kelley Doll, Drew Engelhardt, Brad Gomez, Chris Hare, Kelsey Houser, Bob Jackson, Wolfgang Karlstetter, Holger Kern, David Macdonald, Jessica Parsons, Ron Rapoport, John Barry Ryan and students in his Spring 2018 Public Opinion class.

ABSTRACT
Political surveys often include multi-item scales to measure individual predispositions such as authoritarianism, egalitarianism, or racial resentment. Scholars use these scales to examine group differences in these predispositions, comparing women to men, rich to poor, or Republicans to Democrats. Such research implicitly assumes that, say, Republicans' and Democrats' responses to the egalitarianism scale measure the same construct in the same metric. This research rarely evaluates whether the data possess the characteristics necessary to justify this equivalence assumption. We present a framework to test this assumption and correct scales when it fails to hold. Examining 13 commonly used scales on the 2012 and 2016 ANES, we find widespread violations of the equivalence assumption. These violations often bias the estimated magnitude or direction of theoretically important group differences. These results suggest we must reevaluate what we think we know about the causes and consequences of authoritarianism, egalitarianism, and other predispositions.

KEYWORDS: public opinion; political psychology; survey research; measurement invariance; differential item functioning

Replication files are available in the JOP Data Archive on Dataverse (http://thedata.harvard.edu/dvn/dv/jop).

The research reported here was approved by FSU's Human Subjects Committee (HSC 2017.22647, 2018.22989, and 2018.26523)

To study public opinion and voting is to study human psychology. Recent scholarship has drawn from psychological theories to characterize differences in citizens based on their stable, enduring predispositions. These predispositions include citizens' core values or morals, such as individualism, equality, and fairness (Clifford 2014; Hatemi, Crabtree and Smith 2019; Ryan 2017); their social orientations such as authoritarianism (Hetherington and Weiler 2009; Hetherington and Suhay 2011; Stenner 2005), ethnocentrism (Kam and Kinder 2012; Kinder and Kam 2010), and racial resentment (Banks and Valentino 2012; Kinder and Sanders 1996; Tesler 2012); and their personality traits such as conscientiousness, extraversion, and agreeableness (Gerber et al. 2011, 2012; Mondak and Hibbing 2011). No survey item alone can adequately capture the variation in these broad predispositions. Instead, these predispositions are typically measured with multi-item scales, many of which are included in the American National Elections Study (ANES). Recent research has made heavy use of these scales (e.g., Druckman and Leeper 2012; Federico, Fisher and Deason 2017; Hajnal and Rivera 2014; Hetherington and Suhay 2011; Hetherington and Husser 2012; Hutchings, Walton and Benjamin 2010; Miller, Saunders and Farhart 2016; O'Brien et al. 2013; Sides, Tesler and Vavreck 2019).

Multi-item scales provide great advantages over single-item measures (Ansolabehere, Rodden and Snyder Jr 2008), but their analysis requires assumptions that researchers often overlook. Researchers typically average each individual's responses to the scale items and compare how these averages vary across demographic, social, or political groups. When making these comparisons, researchers assume that the underlying predisposition the scale measures within one group is sufficiently comparable to the underlying predisposition it measures within another group—an assumption known as measurement equivalence or measurement invariance. Without measurement equivalence, the scale cannot provide meaningful group comparisons (Gregorich 2006). Though political scientists rarely evaluate this assumption in survey research, establishing equivalence is akin to the comparability we all seek in our everyday decisions. When deciding between job offers in Boston and Indianapolis, one would not compare the salaries without first adjusting for the cost of living in each city. Without adjustment, the comparison will be misleading because a dollar goes

further in Indianapolis. Likewise, a multi-item scale will mislead when *something other than the underlying predisposition of interest* causes one group to systematically respond differently than another. For example, if voters feel stronger pressure than nonvoters to give socially desirable responses, the Negative Black Stereotypes scale lacks equivalence by turnout because a voter will be expected to receive different scores on the scale than a nonvoter who holds equally strong stereotypes.

Lacking equivalence, between-group comparisons serve no purpose because they compare apples to oranges; one group's values reflect a different concept or are in a different metric than another group's. As a result, analyses that fail to assess the scale's validity often come to the wrong conclusion, misestimating the magnitude or direction of group differences (Abrajano 2015; Pietryka and MacIntosh 2013; Pérez and Hetherington 2014; Stegmueller 2011). This problem plagues a large body of research since these biased estimates of group differences will be reflected in correlation analysis and regression estimates. Despite this problem, political scientists working with multi-item scales rarely check for equivalence.[1] As a result, much of what we *think* we know may be wrong about the distribution of predispositions in the electorate, the causes of these predispositions, and their consequences.

We evaluate the extent of this problem by examining 13 of the most commonly used scales included in both the 2012 and 2016 ANES.[2] Rather than strategically selecting only a few scales or groups to demonstrate our point,[3] we evaluate as many as feasible. We examine which scales lack equivalence for which groups, describe a method to correct scales lacking equivalence, and demonstrate how researchers' conclusions are likely to change when using the corrected scales rather than the uncorrected, off-the-shelf scales. Our analysis suggests that *all* of the uncorrected ANES scales lack measurement equivalence

[1] On 2020-07-16, we found 162 hits when searching Google Scholar for articles including the phrase "american national election study" published since the year 2000 in the *American Journal of Political Science*, *American Political Science Review*, or *Journal of Politics*. This number drops to only three if the search also includes any one of the phrases "measurement equivalence", its synonym "measurement invariance", or a related concept known as "differential item functioning."

[2] The ANES is the most common data source for political behavior research, by one estimate appearing in about a third of published work on the topic between 1980 and 2018 (Robison et al. 2018).

[3] For a discussion of the problems with such approaches, see Harden, Sokhey and Wilson (2019).

for at least some theoretically important groups. We hope the widespread problems we demonstrate encourage survey researchers to adopt new routines to comply with well-established measurement theory. Despite decades of published work about measurement equivalence, the pervasive violations of the equivalence assumption we find make clear that opinion and behavior researchers have paid insufficient attention to this assumption. Scholars may have previously taken an "innocent until proven guilty" approach, assuming their scales are equivalent unless the literature has demonstrated a problem, because previous work has identified equivalence problems only for specific scales and groups. For example, Pérez and Hetherington (2014) demonstrate the authoritarianism scale lacks equivalence for comparing white and black respondents. Our results suggest the burden of proof should be reversed. Scholars risk biased conclusions unless they demonstrate measurement equivalence before conducting their analyses of interest.

Beyond this methodological contribution, our results provide substantive insight into the distribution of predispositions in the mass public. By purging errors induced by inequivalent scales, we offer more accurate estimates of how personality, values, racial attitudes and other predispositions vary with party, gender, and other important groups. The corrected scales often lead to different conclusions than the uncorrected scales would suggest. In some cases, the conclusions differ in magnitude. For example, in 2012 the off-the-shelf scale exaggerates the differences in egalitarian values between rich and poor citizens. In others, the conclusions differ in direction. For instance, the off-the-shelf scale suggests Obama supporters were less authoritarian on average than Romney supporters, but the corrected scale suggests the opposite. As we discuss below, this result may indicate the true relationship between authoritarianism and voting, but may alternatively reflect heretofore unnoticed problems with the ANES Authoritarianism scale's construct validity. Either way, our results highlight the need for greater theoretical development. If the results reflect substantively compelling relationships, they challenge many established theories about these predispositions. If the results reflect poor construct validity, we must reevaluate extent theories because so much of their empirical verification rests on these scales.

A THEORY OF MEASUREMENT AND BIAS

When we compare different groups using a multi-item scale, we must assume that the items exhibit measurement equivalence, capturing the same construct for each group. Multi-item scales may lack equivalence, however, because the ways people interpret and respond to questions often differ systematically between social groups. All survey questions and response options contain ambiguity. Consequently, some respondents will interpret even a carefully worded item differently than will other respondents. Respondents' personal backgrounds shape their interpretations, causing their understanding to differ from individuals with dissimilar educations, ethnicities, or other social circumstances. For instance, the Authoritarianism scale asks respondents to choose which of two desirable traits is more important for a child to have. One item asks whether it is better for a child to be considerate or well behaved. This item promotes inequivalence by gender if women differ from men in their conception of a "well behaved" child.

Scales also lack equivalence when response biases vary from group to group. For instance, some groups may feel more compelled than others to provide socially desirable responses (Ansolabehere and Hersh 2012). Education, in particular, predicts many different response biases (Narayan and Krosnick 1996) such as acquiescence—the tendency to choose more agreeable response options to agree/disagree items (but see Lelkes and Weiss 2015). These biases are likely to produce inequivalence, confounding estimates of group differences. For instance, the Egalitarianism scale relies on questions with agree or disagree anchors, and thus we should expect less-educated respondents to choose more agreeable options than better-educated, but similarly egalitarian individuals. Since education covaries with many important grouping variables, measurement equivalence may be rare without correction.

Though rarely invoking the technical term "measurement equivalence," scholars have criticized various multi-item scales for failing to meet this standard. Consider the Racial Resentment scale (alternatively labeled symbolic or modern racism), designed to measure white respondents' views about whether African Americans deserve special government assistance (Kinder and Sanders 1996). Some scholars have criticized this scale for conflating racial animus with policy preferences (Carmines, Sniderman and Easter 2011; Feldman

and Huddy 2005) or political sophistication. Gomez and Wilson (2006) argue that less sophisticated individuals are less likely to attribute individual outcomes to systemic forces, and hence they are less likely to link racial inequality to systemic causes or solutions. By implication, the Racial Resentment scale will provide a biased estimate of the relationship between racial resentment and policy preferences or sophistication. From a measurement perspective, the Racial Resentment scale is not sufficiently unidimensional because the policy preference and sophistication constructs have intruded. Moreover, comparisons of other groups will lack validity if the groups differ in average preferences or sophistication levels. If college graduates tend to hold different policy preferences than non-graduates, then the estimated education gap in racial resentment may arise from actual differences in racial animus or qualitatively distinct differences in policy preferences.[4]

A well-developed framework exists to test for the presence of measurement equivalence (Andrich 2013; Bond and Fox 2015; Gregorich 2006; Rasch 1980). The political knowledge literature provides a rare example where political scientists have applied this framework, finding that the apparent gender gap in political knowledge arises in part as an artifact because knowledge scales lack measurement equivalence by gender (Lizotte and Sidman 2009). When unsure about the correct answer, men tend to guess more frequently than equally knowledgeable women and, consequently, average higher scores (Mondak and Anderson 2004). Similarly, political knowledge scales lack equivalence for comparisons by age, education, income, race, and turnout (Abrajano 2015; Pietryka and MacIntosh 2013). Scholars have found inequivalence problems for other scales when comparing responses administered in different languages (Pérez 2011) or countries (Stegmueller 2011). And Hare et al. (2015) demonstrate that the U.S. electorate appears considerably more polarized once measurement equivalence has been established. In summary, the scales and grouping variables that have received systematic analysis often lack measurement equivalence. Despite this work, most survey research relying on multi-item scales assumes equivalence, but fails to check this assumption. We seek to assess how problematic that omission might be.

[4]See (Engelhardt 2020) for a discussion of the racial resentment scale's measurement equivalence by age.

Meaningful comparisons require measurement equivalence

To establish measurement equivalence, we follow the procedure summarized in Figure 1 and explained in the following sections.

To develop the formal approach for evaluating measurement equivalence, imagine we wish to examine the relationship between gender and support for limiting the government's role in domestic affairs. In this case, we could examine the three dichotomous items that form the ANES Limited Government scale. Typically, researchers sum or average an individual's responses to a scale's items, using the averages as marks along a latent continuum enabling the placement of respondents in relation to each other. This form of concatenation requires a unit of length that consistently iterates in successive segments (Bond and Fox 2015). The observed distance between scores is meaningful, therefore, if the measuring device operates consistently across groups along the latent continuum. We can only compare how strongly women and men value limited government if the three ANES items exhibit comparable measurement properties across these groups. Otherwise, any group differences we observe may be an artifact of an inconsistent measuring device (Gregorich 2006).

An insufficient level of equivalence indicates the data cannot support meaningful comparisons because of noise in the measurement. In some instances the group differences observed are due to this noise, but are mistaken as actual differences on the latent trait. For instance, we may mistakenly attribute a gender gap on the Limited Government scale to real differences in the extent to which women and men value limited government when instead the gap arises from differences in response patterns irrelevant to support for limited government. In other cases, the measurement noise obscures actual trait differences, perhaps making women seem more similar to men than they really are. Either way, the threat of inequivalence casts doubt on the study's internal validity. This threat can be removed by establishing sufficient measurement equivalence across the most important covariates.

Differential item functioning indicates inequivalence

To evaluate this threat, we can test for differential item functioning (DIF), which indicates that an item from the scale operates inconsistently across groups and therefore lacks measurement

equivalence. In the case of limited government, an item shows DIF by gender when a woman is expected to give different responses than a man with equal preferences for limited government. When researchers average the items in a scale such as Limited Government, they assume the score reflects a single dimension (Jacoby 1991, 40), capturing only the latent trait of interest. DIF arises from unintended multidimensionality (Ackerman 1992) in which some "nuisance" dimension is distributed unequally between subgroups. As discussed above, the nuisance dimension can reflect many factors such as salience or prior socialization. This nuisance dimension intrudes on the measurement occasion, creating a *group* × *item* interaction that is observed after controlling for the trait of interest. If we can identify items with DIF for our comparisons of interest, we can take corrective action to derive a unidimensional measuring device that *is* sufficiently equivalent to fulfill its intended purpose.

One means of identifying DIF is to assess how well the data conform to the Rasch (1980) model. The Rasch model represents a platonic form of fundamental measurement (Wright 1999), providing interval-level measures which form the basis for regression analysis and other common statistical comparisons. As no real-world data can be expected to fit a platonic model exactly, our interest lies primarily in the critical ways in which the data may fail to fit.

The Rasch model provides a clear, direct, and easy to interpret indication of differential item functioning thanks to a property known as *specific objectivity*. This property means that, when the data approximately fit the model, comparisons between any two persons are independent of which items are selected from a class of items that are designed to measure the construct (Andrich 2004). Likewise, comparisons of items are independent of which persons participated in the survey. For dichotomous items, the (natural) log probability of endorsement versus non-endorsement is the difference between the relative locations on the latent continuum of item i (D_i), and survey participant n, (B_n):

$$ln(P_{ni1}/P_{ni0}) = B_n - D_i \tag{1}$$

For example, the first item on the Limited Government scale asks, "Which of the two statements comes closer to your view? A) The main reason government has become bigger over the years is because it has gotten involved in things that people should do for themselves.

B) Government has become bigger because the problems we face have become bigger." A respondent is said to endorse limited government if they choose option A. The Rasch model places this item and each respondent on the same latent continuum. When a respondent's location equals the item location, their probability of endorsing the item equals 50 percent. More generally, this probability increases as the respondent's location increases relative to the item location.

For polytomous items, which have three or more response categories, a term (F_{ij}) is added for the $j = 0, 1, \ldots m$ thresholds between categories to derive the Rasch partial credit model:

$$ln(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_{ij} \tag{2}$$

Rasch model fit is assessed using standardized residuals between the observed responses and the expected responses predicted by the model. Extensive misfit of the data to the model indicates that additional corrective action must be taken to construct the latent variable (Andrich 2004, 2013). This approach is intended to yield measures that conform as closely as possible to the characteristics of the Rasch model. This approach is in contrast to searching for a model with a sufficient number of parameters to describe the data at the cost of violating fundamental measurement principles.

This study focuses on a fundamental violation of the Rasch model. DIF occurs when an item's estimated location on the latent continuum varies systematically between groups. This variation produces group differences in the standardized residuals. Therefore, detecting uniform DIF for any item requires only a simple one-way ANOVA of these residuals based on group membership (Hagquist and Andrich 2004).[5] Since work that has not evaluated measurement equivalence assumes that no DIF exists, we use this assumption as our null

[5]Alternatively, researchers can assess non-uniform DIF using two-way ANOVAs (Hagquist and Andrich 2004) by dividing the latent continuum into "classes" with roughly equal numbers of survey participants and test for *class* × *group* standardized residual mean differences. We forgo this approach because the ANES scales are typically too short to divide into more than four classes, presenting serious violations of the ANOVA assumptions. Examining only uniform DIF biases conclusions *against* finding DIF.

hypothesis, correcting DIF if $p < .05$, after adjusting for multiple comparisons using the Benjamini and Hochberg (1995) method.[6]

The presence of DIF may suggest substantively interesting differences between the groups (Andrich and Hagquist 2015). Even if these differences are substantively interesting, DIF indicates that the items are not measuring a single latent construct as intended. And thus analysis of the uncorrected scale will conflate variation on the trait of interest with one or more extraneous factors. We expand on this point in the conclusion, but stress now that avoiding spurious group differences requires that DIF is eliminated.

Correcting DIF by resolving items

To establish equivalence, all scale items must be evaluated for DIF, correcting it when it is detected. We use the sequential approach recommended by Andrich and Hagquist (2012, 2015) in which each item in the scale is checked for DIF. When one or more items show DIF, the item with the largest significant F-statistic is corrected and the remaining items are then re-checked. DIF is corrected by *resolution*, substituting the original biased item for new group-specific pseudo-items.[7] One new pseudo-item is created for each group, retaining the original item's responses for a specific group, but recoding the responses from other groups as (structurally) missing. The Rasch model is then re-estimated with the pseudo-items acting as separate items with different locations. A new ANOVA is then fit to check for DIF and the process is repeated until no item shows DIF.[8]

Resolving items with DIF, rather than the common alternative of omitting them from the scale, offers several benefits (Andrich and Hagquist 2012). Resolving rather than omitting items with DIF is particularly beneficial for the short scales in the ANES because it retains

[6] If the greater concern is failing to correct DIF, one could choose a higher p-value or avoid statistical significance thresholds entirely. Correcting an item with no real DIF using the method we describe below will result in virtually identical location estimates (Masri and Andrich 2020, 176).

[7] This correction is known as *resolution* because it is analogous to a chemical resolution, in which a mixture is separated into its constituent parts.

[8] The difference in the locations of the pseudo-items represents the magnitude of the DIF, measured in the same metric as the latent trait. After all items have been resolved, the DIF summed over all items indicates the magnitude of bias that this resolution procedure has removed, again measured in the same metric as the latent trait.

greater reliability and provides finer distinctions between individuals at difference positions along the latent trait (Bakker and Lelkes 2018; Hagquist and Andrich 2017). These more precise estimates provide greater statistical power to detect DIF. Further, resolving DIF sequentially, rather than correcting all items demonstrating DIF in a single step, avoids the "artificial" DIF that misleadingly appears to offset the bias created by items with real DIF. Failing to address artificial DIF, researchers often mistakenly conclude that the scale-level DIF (DIF accumulated across all items, sometimes labeled 'Differential Test Functioning') appears negligible (Andrich and Hagquist 2012, 2015). By retaining the items exhibiting only artificial DIF, sequential resolution also offers greater content validity and statistical power through greater reliability and lesser measurement error (see Masri and Andrich 2020).

Complete resolution is not always possible because valid group comparisons require at least one DIF-free item. Since the resolved pseudo-items are group specific, a DIF-free item is required as an anchor, establishing the latent trait's origin and placing the groups in the same metric. To compare how far your salary will go in Boston relative to Indianapolis, you might compare the price of some good such as a pair of shoes. This comparison requires that identical shoes are sold in both cities. Likewise, comparing scale values of women and men requires an anchor item that operates equally for both groups. Anchors may be difficult to obtain in the ANES data, however, because most scales feature four or fewer items. When complete resolution is not possible, one may elect to use a partially corrected scale that resolves all but the last item showing DIF. The partially corrected scale reduces, but does not eliminate bias from DIF. Therefore, between-group comparisons are *less* biased with these scales than with the off-the-shelf scales. But partially corrected scales nonetheless lack optimal measurement equivalence. When DIF cannot be completely resolved, researchers should note this limitation and, when possible, develop alternative measures to capture all relevant variation.

The short scales on the ANES also limit their content validity, reliability, and statistical power to detect DIF. With limited statistical power, violations of the measurement equivalence assumption may go unnoticed. If researchers wish to provide evidence *against* DIF, they

may instead use their substantive expertise to determine the negligible range of test-level DIF.[9] If the confidence interval for test-level DIF falls entirely within this negligible range, it provides evidence against the null that DIF exists. For an implementation of this method using an alternative DIF test, see Casabianca and Lewis (2018).

Several alternative approaches exist to evaluating and establishing measurement equivalence. If researchers anticipate DIF, they may include anchoring vignettes on their survey batteries (King et al. 2003). When data without vignettes have already been collected, they may rely instead on multi-group confirmatory factor analysis or Mokken (1971) scaling. We discuss these and other alternatives in the online Supporting Information (SI) section A. While we believe the Rasch approach offers some advantages for this application over alternative methods, the more important distinction is that any reasonable method for evaluating the measurement equivalence assumption constitutes an improvement over the common alternative of leaving the assumption implicit and unscrutinized.

DATA: 2012 AND 2016 ANES

We offer a representative assessment of the problems DIF creates by examining 13 prominent multi-item scales included in the American National Election Studies. Researchers seeking to demonstrate a methodological point often choose a few published articles to re-analyze, demonstrating how different methods yield different substantive conclusions. This approach often overstates the importance of the methodological point in question because the studies selected for replication are not representative. Rather, researchers tend to replicate studies that yield the greatest support for their proposed methods (Harden, Sokhey and Wilson 2019). Instead of replicating a handful of published studies, we therefore examine almost all[10] of the commonly used scales that are available in identical formats in both the 2012 and 2016 American National Election Studies. And we likewise examine the grouping

[9]Test level DIF can be measured as the between-group difference in item locations after resolution, summed over all scale items.

[10]We omit the political knowledge battery because its measurement properties have already received considerable scholarly attention (e.g., Abrajano 2015; Lizotte and Sidman 2009; Pietryka and MacIntosh 2013). We omit the internal and external efficacy scales because the 2012 study randomly assigned respondents to receive one of two sets of questions. The 2016 ANES included two items from one of those

variables we commonly see in analyses of these scales. Further, we conducted exploratory analysis using the 2012 data and, based on the results, preregistered our analysis of the 2016 data—before the 2016 data were released.[11] The preregistration ensures that we report all results rather than just those from the scales or groups that support our argument.

The scales: We examine the following scales: Authoritarianism, Egalitarianism, Limited Government, Moral Traditionalism, Negative Black Stereotypes, Non-Voting Participation, Racial Resentment, Wordsum, and the five personality traits from the Ten Item Personality Index, or TIPI (Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness To Experiences). We describe these scales and explain their construction in SI-B1.

The grouping variables: We examine whether these scales exhibit equivalence for ten grouping variables: gender, party identification, liberal-conservative ideology, electoral turnout, race/ethnicity, education level, age, income, presidential vote choice, and survey mode. Aside from survey mode, we chose these variables for their theoretical importance across a broad range of opinion and behavior research. We chose survey mode for its methodological importance (Malhotra and Krosnick 2007). The ANES, previously conducted entirely through face-to-face interviews, now relies on interviews conducted either face-to-face or over the internet. By examining DIF across survey modes, we test the implicit assumption that these two sets of responses are comparable.

There are, in principle, no restrictions on how groups can be operationalized. For example, the joint effect of a three-category race variable and binary gender can be assessed by creating six race × gender groups. In practice, however, using many groups will reduce the subsample sizes, limiting precision and statistical power. Though the Rasch model is better able to handle small subsamples than alternative approaches such as confirmatory factor analysis or more complex item response models (Belzak 2019), the statistical power to detect DIF in the Rasch decreases as group sizes become smaller and less balanced in

sets and two from the other, and thus no 2012 respondents received the same scales as any 2016 respondents.

[11] The preregistration was completed on 2017-03-24, prior to the 2017-03-31 release of the 2016 ANES data. The preregistration form can be found at `https://osf.io/9n4zs/`

size.[12] DIF estimates will be less reliable when we use national samples to examine relatively uncommon groups such as Native Americans, Libertarians, or transgender people. We therefore focus our analysis on relatively common subgroups, which we describe in detail in SI-B2.

EMPIRICAL RESULTS

An analyst relying on multi-item scales should first test for measurement equivalence among the groups of greatest theoretical importance. If DIF is found for one or more items, the analyst must correct the DIF if possible. If the DIF can be corrected, the analyst may then use the corrected scale to examine group differences. We demonstrate each of these steps.

Testing for DIF

As a first step, we fit a Rasch model to each scale and then examine whether the results exhibit DIF for each grouping variable. We then examine DIF with the one-way ANOVAs described above and in more detail in SI-C. As an example, we highlight DIF for the Egalitarianism scale grouped by education and income. Figure 2 shows the ANOVA p-values for each item. The first panel shows that in 2012, items 1, 3, and 4 exhibit significant DIF by education. This result suggests that individuals with the *same* latent level of egalitarianism have *different* expected responses for these items, depending on the extent of their education. The scale exhibits similar problems for education in 2016 and income in both years. Therefore, differences in scores on the scale across the range of education or income do not necessarily indicate real differences in egalitarianism. Nonetheless, we can correct this problem for education in 2012 and income in both 2012 and 2016. We cannot correct this problem for education in 2016, however, because all the items exhibit DIF.

Figure 3 summarizes the analogous DIF tests for all scales and grouping variables.[13] In the figure, dark boxes indicate that the item exhibited DIF for that grouping variable. When all items exhibit DIF, the scale cannot be corrected and is therefore invalid, as indicated by an X to the left of the items. Many of the items show DIF. In total, 58% of item-grouping

[12] Simulation studies suggest that the Rasch provides adequate power to detect uniform DIF with fewer than 500 observations, even for scales with fewer than six items (see Belzak 2019; Scott et al. 2009).

[13] The full ANOVA results can be found in Tables C1–C13 of SI-C.

variable combinations showed DIF in both years, 26% showed DIF in one of the two years, and only 17% did not show DIF in either year. The items showing DIF in 2012 also tend to show significant DIF in 2016. Among the item-grouping variable combinations showing DIF in 2012, 76% of the cases show DIF in 2016, compared to 32% of the cases showing no DIF in 2012. These results provide preliminary evidence that researchers may be misled if they fail to check for DIF. Since statistically significant DIF does not necessarily indicate substantively important bias, however, we examine below the extent to which DIF affects the conclusions researchers might draw from these scales.

Despite the general continuity of the results from 2012 to 2016, several differences emerge. Some differences are likely the result of low scale reliability; the brevity of the ANES scales produces high levels of measurement error (Ansolabehere, Rodden and Snyder Jr 2008). These differences may also develop over time for substantive reasons such as evolving patterns of elite rhetoric. For instance, both items on the Negative Black Stereotypes scale exhibited DIF by vote choice in 2012, but neither item did so in 2016. Donald Trump's use of explicitly racist campaign appeals, in contrast to the more implicit appeals over the last several decades, may serve as a partial explanation for this change. Trump's rhetoric may have made racist voters in 2016 more willing than in 2012 to admit their racist beliefs (Valentino, Neuner and Vandenbroek 2018), causing DIF by vote choice to weaken.

The analysis lends additional support to previous work examining individual scales and grouping variables. Our authoritarianism results reinforce Pérez and Hetherington (2014) who find that the Authoritarianism scale lacks equivalence between black and white respondents. Figure 3 suggests this problem extends to many other grouping variables such as party identification and education.[14] Though the other scales we examine have not previously received formal tests for measurement equivalence, the results are consistent with the work arguing that the Racial Resentment scale conflates policy views with racial animus (Carmines, Sniderman and Easter 2011; Feldman and Huddy 2005). The Racial Resentment items all show DIF for ideology, party identification, and vote choice in both years. Further work is needed to determine whether policy views are the source of this bias.

[14]In exploratory analysis, we find similar results when the data are restricted to non-Hispanic whites.

Though all scales exhibit DIF, some have more problems than others. On one end of the spectrum, all items from the Limited Government scale exhibit DIF for all grouping variables in 2012 and all but one in 2016. On the other end of the spectrum, the Negative Black Stereotypes items exhibit DIF in only a few instances. Likewise, some grouping variables show DIF for most items. This analysis reveals that people differentiated on these traits, which include ideology, party ID, and education, interpret and process what may appear superficially to be the same political stimuli in fundamentally different ways—almost as if they live in different political realities. These interpretations are so different, unfortunately, that without correction they may confound comparisons on the traits of theoretical interest. Other grouping variables, such as gender, show DIF for relatively few items.

The widespread item-level validity problems can nonetheless be corrected for many grouping variables. This correction is possible as long as one or more items lack DIF, providing an anchor linking the groups. By this criterion, correction is possible in both years for 38% of the scale-grouping variable combinations, and 27% in one of the two years. Still, 35% of the scale-grouping variable combinations cannot be corrected in either year. For scales that cannot be corrected, the data suggest the groups differ qualitatively to the point that they are not quantitatively comparable with the existing items. For example, the data reveal that party identification groups differ on moral traditionalism to the extent that they cannot be considered subgroups from the same population, at least as the construct is defined using this scale. This DIF is *not* evidence that the parties differ in mean levels of moral traditionalism. Rather, this DIF indicates the data generated by the items used to measure moral traditionalism cannot be used to make such comparisons. Similar observations hold for other constructs, such as support for limited government, authoritarianism, and egalitarianism. In these cases, the DIF cannot be corrected as the scales are not sufficiently unidimensional and are not consistent measuring devices across groups.

Correcting DIF

Figure 3 shows that almost all scales have at least one item showing DIF for almost every grouping variable. We therefore apply the sequential correction explained above to all items showing DIF and estimate each person's location on the latent trait from the final

Rasch model. These corrected scales provide valid group comparisons on the latent trait if they satisfy two conditions. First, the final model must reveal no remaining DIF. Second, the items that remain unresolved capture sufficient variation in the latent trait to satisfy the researcher's needs. The first condition can be evaluated empirically, but the second ultimately requires the researcher to make a subjective call about the measure's content validity.

As an example, Figure 4A shows how the 2012 Egalitarianism scale can be corrected for DIF by education and income. The dashed line plots each item's baseline location estimate before DIF was corrected. Recall from Equation 2 that an item's location indicates how egalitarian someone would need to be in order to have an equal chance of responding above or below the midpoint of the item. The greater the item's location, the lower the likelihood of choosing an egalitarian response. The solid lines indicate how the location varies with education and income after DIF has been corrected.

Consider education, in the left panel of Figure 4A. The first condition required for correction is met because item 2 exhibits no DIF and thus its location is comparable for people of all education levels. With this fixed location, the other items' locations can vary with education, but their relative distance from item 2 provides a means to keep them in a comparable metric. To satisfy the second condition, the researcher must decide whether item 2 sufficiently captures egalitarianism. If so, they may proceed with the analysis. If not, they must decide whether they prefer the uncorrected measure that is confounded by DIF, or the corrected measure that captures the construct of interest only incompletely. Whichever measure they choose, they should report sensitivity analysis examining how this choice affects their results, as we demonstrate in the next section.

If we believe item 2 provides sufficient content validity, we can use the item locations to learn more about the potential sources of DIF. The left panel of Figure 4A shows that the corrected locations of items 1 and 4 increase with education. This pattern suggests that better educated people are less likely to choose egalitarian responses for these items than are people who are equally egalitarian, but less educated. Holding their true levels of egalitarianism constant, then, better educated people will tend to receive lower scores

on the off-the-shelf scale. In the right panel, items 1 and 4 exhibit similar problems for income. Note that egalitarian responses for both items require agreement with the prompt, in contrast to items 2 and 3 which require disagreement. Thus the DIF may stem in part from acquiescence bias which tends to decrease with socioeconomic status (Narayan and Krosnick 1996; Rammstedt, Danner and Bosnjak 2017). Of course, this conjecture is only speculative and it is beyond the scope of this study to determine the root causes of the DIF. The evidence makes clear, nonetheless, that these items do not operate as intended.

The Consequences of DIF

Working with large samples like those found in the ANES can lead to statistically significant DIF even when the DIF produces negligible substantive impact on the comparisons of interest. Thus, before abandoning the off-the-shelf scale, researchers should evaluate whether the magnitude of DIF in the data leads to substantively different conclusions than the corrected scale. If scores for the off-the-shelf Egalitarianism scale conflate egalitarianism with education and income, then estimates of the relationships between these variables may be biased. To demonstrate this point, Figure 4B plots the Egalitarianism scale's estimated bivariate relationships with education and income. Each panel in the figure provides coefficients from two linear regressions. In the first, the off-the-shelf, uncorrected scale is regressed on a grouping variable. In the second, the corrected, DIF-free scale is regressed on the same grouping variable.[15] Since both grouping variables are ordinal, we include a dummy for each value, omitting the minimum value as the reference category.

Figure 4B suggests that the off-the-shelf Egalitarianism scale produces biased estimates of the latent trait's relationships with education (left panel) and income (right panel). For example, the off-the-shelf scale suggests college graduates are significantly less egalitarian than those without high school degrees. In contrast, the corrected scale suggests a weak, insignificant relationship in the *opposite* direction. The difference in these estimates is

[15] We standardize the off-the-shelf scores and corrected scores with mean = 0 and standard deviation = 1. We estimate the regressions with Taylor series standard errors (DeBell 2010)

roughly a quarter of a standard deviation in egalitarianism.[16] Likewise, the off-the-shelf scale suggests the richest income group tends to be much less egalitarian than the poorest group. The corrected scale suggests a significantly weaker relationship.[17] In summary, scholars relying on the uncorrected scale may incorrectly conclude that a strong, negative relationship exists between egalitarianism and these indicators of socioeconomic status.

These large substantive differences emerge despite strong correlations between the off-the-shelf and corrected scales. In 2012, the off-the-shelf Egalitarianism scale has a .98 correlation with the scale corrected for education and a .97 correlation with the income-corrected version. These strong correlations occur among all of the off-the-shelf scales and their corrected counterparts, as shown in Figure E1 in SI-E. The correlations all exceed 0.8 in 2012 and 0.9 in 2016. These strong correlations suggest the corrected scales retain most of their content after purging them of DIF-induced bias.

Despite these strong positive correlations, the off-the-shelf scales often suggest different substantive conclusions than do their corrected counterparts. Figure 5 displays the estimated relationship between each scale and each grouping variable.[18] The figure is restricted to scale-group combinations that lacked significant DIF for at least one item, enabling correction. For exploratory analysis of the scales showing DIF for all items, see Figure F1 in SI-F.

The results in Figure 5 should give pause to researchers working with off-the-shelf scales. While many relationships remain unchanged, a large proportion of the corrected estimates differ significantly from the uncorrected ones. In these cases, the off-the-shelf scale will lead to biased conclusions because not all items are operating as intended. For instance, the off-the-shelf 2012 Negative Black Stereotypes scale exaggerates the differences between Republican and Democratic respondents. In many cases, the direction of the relationship reverses. For example, the 2012 off-the-shelf Authoritarianism scale suggests that authoritarians tend to vote Republican, but the corrected scale suggests the reverse.

[16]The difference between the off-the-shelf and corrected coefficients is -0.22 (95% Confidence Interval = [-0.42, -0.03]). To measure uncertainty in the difference between the model estimates, we use pooled-sample standard errors.

[17]The difference between the estimates is -0.18 (95%CI = [-0.31, -0.06]).

[18]These estimates are derived from the same process described for Figure 4B.

Likewise, the uncorrected Non-Voting Participation scale suggests women participated less than men in the 2012 election, but the corrected scale suggests women participated more than men. Survey mode seems to be a particularly problematic grouping variable for many scales, suggesting that adding a mode indicator in a regression cannot control for differences between modes. Given the prevalent DIF, the model is unlikely to capture adequately the covariation between the mode indicator and the latent trait of interest.

Many of the corrected results differ from previous research or theoretically grounded expectations. The widespread presence of DIF may suggest problems with the underlying theory, but it may instead suggest problems with the items used to measure the constructs of interest. These measurement problems may have gone unnoticed in previous research because the off-the-shelf scales happened to produce the expected relationships with criterion variables. To avoid this problem, tests of construct validity must come after differential item functioning has been eliminated. The broad differences between these results and past work provide an opportunity for future scholars to reexamine past findings after removing bias caused by measurement inequivalence.

DISCUSSION

Readers who lack familiarity with psychometric models may be puzzled when empirical relationships between the scale and grouping variable are directly opposed to what may have been anticipated by the survey designers. The 2012 Authoritarianism scale provides a notable example, where Republican voters average lower scores on the latent variable than Democratic voters. This pattern emerges because three of the four items operate differently for the two groups, as demonstrated in Figure 3. This DIF does *not* indicate differences between Republican and Democratic voters on the latent trait measured by the scale, but rather that Republican voters were influenced by different combinations of extraneous factors than Democratic voters, causing the survey questions to function dissimilarly. The (unidentified) nuisance dimension(s) makes it easier for Republican voters to endorse three of the child rearing items, relative to Democratic voters with the same

off-the-shelf score.[19] In other words, Democratic voters need to be more authoritarian than Romney voters to endorse the authoritarian option on these items. Consequently, the easier "items" Republicans respond to result in a lower estimated group mean on the latent variable as shown in Figure 6. The two groups are effectively answering different questions and the off-the-shelf scores are thus not comparable. We elaborate on this point in SI-G.

We are not arguing that we have found the true relationship between vote choice in 2012 and authoritarianism. Rather, we argue that these data provide empirical evidence of an unbiased relationship between vote choice and the latent trait captured by the ANES Authoritarianism *scale*. We hope future work will examine whether this result reflects a compelling substantive relationship or whether it indicates that the scale lacks construct validity. The challenge will be to avoid returning to a tautological measure. Researchers first introduced these items, which focus on child-rearing preferences, because alternative measures of authoritarianism included items about political preferences, conflating authoritarianism with its theorized consequences (Stenner 2005). If scholars reject the child-rearing scale based solely on its correlation with political preferences, they will negate this benefit.

CONCLUSION

Political scientists have placed an increasing premium on causal inference, but our analysis demonstrates an overlooked threat for many recent causal claims. Measurement equivalence is a fundamental element of internal validity. Its absence leaves any study vulnerable to the justified criticism that demonstrated effects (or lack thereof) may be an artifact of poorly constructed measuring devices. As noted above, this consideration is particularly relevant when the observed effects are small in magnitude. Therefore, the analysis we present here holds important implications both for scholars analyzing data already collected and for those designing new survey batteries.

For scholars relying on previously collected data, our results suggest they must examine whether the equivalence assumption holds for their scales. Failing to do so, they stand a substantial chance of reaching biased conclusions because some items may not be sufficiently

[19] With the Rasch model's specific objectivity, the off-the-shelf score is a sufficient statistic to estimate the latent trait score.

equivalent across groups and, therefore, the data they generate will not support meaningful comparisons. All of the 13 scales we examine lack measurement equivalence for theoretically important grouping variables (Figure 3). For instance, each scale includes items exhibiting DIF by partisanship and ideology in 2012. We find similar results for these groups in 2016 and for other grouping variables in both years. These results suggest a number of unidentified dimensions are unequally distributed between groups. These unidentified dimensions pose a nuisance when the trait is assumed to capture only a single dimension, but they may reflect substantively interesting group differences (Andrich and Hagquist 2015).

A key challenge for future work will be separating the DIF that represents measurement artifacts from that which represents theoretically relevant group differences. Even if the DIF is theoretically relevant, its presence indicates that these differences are manifesting in qualitatively different ways, and thus the uncorrected scale will conflate variation measured in different metrics. In such cases, researchers should strive where possible to develop an additional scale or scales to capture all relevant dimensions. Unfortunately, identifying substantively relevant DIF is more challenging for the study of political attitudes than it is for other disciplines. In education research, where item response theory developed, assessing the relevance of DIF is more straightforward because students are typically tested soon after their opportunity to learn the material. In contrast, political scientists often study political attitudes that may be measured years after their suspected causes, introducing great complexity into the causal chain. When many explanatory variables are themselves causally related to one another, as is the case with social identities and political attitudes, it is difficult to discern the root causes of DIF and therefore difficult to determine the substantive meaning of the DIF. This challenge is by no means insurmountable. Pérez and Hetherington (2014) offer a model for such work, drawing evidence from both observational and experimental designs to understand the effects of race on DIF in the child-rearing authoritarianism scale.

Given the widespread inequivalence we detect, we present a simple method to improve measurement using the data at hand. Though the off-the-shelf scales lack equivalence, this method resolves the problem for many scale *by* grouping variable combinations. Using a strict definition of DIF, all items are retained, albeit some in a form not intended by

the designers, and DIF is eliminated to the extent possible. In addition, we identify scale *by* grouping variable combinations that cannot be resolved. When correction is possible, the relationship between many scales and grouping variables changes in magnitude or direction from the ones produced by the uncorrected—and unvalidated—off-the-shelf scales (Figure 5). Therefore, scholars should not assume results are valid unless measurement equivalence has been established. This point holds even if the scale is used only as a control, rather than as a key outcome or explanatory variable. If the off-the-shelf scale misestimates the relationship between the latent trait and *either* the outcome variable or the explanatory variable of interest, including it as a control will fail to eliminate the bias its introduction was intended to address.

For those designing new surveys, our results suggest that scholars should include as many items per scale as they can. Many of the items we examine exhibit DIF for important groups, but some items exhibit DIF for some groups while lacking DIF for others. Including more items therefore increases the likelihood that researchers can construct a valid scale for the groups they are interested in comparing. Longer scales also allow finer distinctions between levels on the latent trait (Bakker and Lelkes 2018). Since survey space is limited, survey designers must consider the tradeoff between the number of items per scale and the number of different scales they include. If increasing the number of scales decreases the number of items used to measure each one, then including too many scales will limit researchers' ability to use any of them.

We recognize that testing for measurement equivalence adds additional complexity to survey research. Just as political scientists regularly demonstrate the reliability of their scales with Cronbach's alpha coefficients, so too should they demonstrate the validity of their scales with measurement equivalence tests. Some scholars may seek to avoid this complexity by relying instead on single-item measures. Yet relying on single-item measures only obscures the problem of measurement equivalence because it cannot be tested—and therefore its violations cannot be observed or corrected.

To understand presidential election outcomes, race relations, policy reform, and other important issues, scholars will continue to mine citizens' psychological predispositions for

explanations. Yet our results suggest that we must reassess much of what we thought we knew about these topics. To gain traction on these issues—and to avoid propagating theories built on noise—researchers, journal editors, and reviewers must place a premium on obtaining data that will permit valid inferences. By seeking measurement equivalence, this task becomes easier, not harder.

References

Abrajano, Marisa. 2015. "Reexamining the "Racial Gap" in Political Knowledge." *The Journal of Politics* 77(1):44–54.

Ackerman, Terry A. 1992. "A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective." *Journal of Educational Measurement* 29(1):67–91.

Andrich, David. 2004. "Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms?" *Medical Care* 42(1):I7–I16.

Andrich, David. 2013. "The Legacies of R. A. Fisher and K. Pearson in the Application of the Polytomous Rasch Model for Assessing the Empirical Ordering of Categories." *Educational and Psychological Measurement* 73(4):553–580.

Andrich, David and Curt Hagquist. 2012. "Real and Artificial Differential Item Functioning." *Journal of Educational and Behavioral Statistics* 37(3):387–416.

Andrich, David and Curt Hagquist. 2015. "Real and Artificial Differential Item Functioning in Polytomous Items." *Educational and Psychological Measurement* 75(2):185–207.

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437–459.

Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder Jr. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(02):215–232.

Bakker, Bert N. and Yphtach Lelkes. 2018. "Selling Ourselves Short? How Abbreviated Measures of Personality Change the Way We Think about Personality and Politics." *The Journal of Politics* 80(4):1311–1325.

Banks, Antoine J. and Nicholas A. Valentino. 2012. "Emotional Substrates of White Racial Attitudes." *American Journal of Political Science* 56(2):286–297.

Belzak, William C. M. 2019. "Testing Differential Item Functioning in Small Samples." *Multivariate Behavioral Research* pp. 1–26.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Bond, Trevor and Christine M. Fox. 2015. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition*. Third ed. New York: Routledge.

Carmines, Edward G., Paul M. Sniderman and Beth C. Easter. 2011. "On the Meaning, Measurement, and Implications of Racial Resentment." *The ANNALS of the American Academy of Political and Social Science* 634(1):98–116.

Casabianca, Jodi M. and Charles Lewis. 2018. "Statistical Equivalence Testing Approaches for Mantel–Haenszel DIF Analysis." *Journal of Educational and Behavioral Statistics* 43(4):407–439.

Clifford, Scott. 2014. "Linking Issue Stances and Trait Inferences: A Theory of Moral Exemplification." *The Journal of Politics* 76(3):698–710.

DeBell, Matthew. 2010. How to Analyze ANES Survey Data. Technical Report nes012492 ANES Technical Report.

Druckman, James N. and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4):875–896.

Engelhardt, Andrew M. 2020. "Generational Persistence in the Nature of White Racial Attitudes." *Working Paper accessed 9/16/2020 at https://drive.google.com/file/d/10N0NWZ7hDFCr0-fyDE5cWdhp6bCu0j1b/view?usp=sharing* .

Federico, Christopher M., Emily L. Fisher and Grace Deason. 2017. "The Authoritarian Left Withdraws from Politics: Ideological Asymmetry in the Relationship between Authoritarianism and Political Engagement." *The Journal of Politics* 79(3):1010–1023.

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.

Gerber, Alan S., Gregory A. Huber, David Doherty and Conor M. Dowling. 2012. "Disagreement and the Avoidance of Political Discussion: Aggregate Relationships and Differences across Personality Traits." *American Journal of Political Science* 56(4):849–874.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, Connor Raso and Shang E. Ha. 2011. "Personality Traits and Participation in Political Processes." *The Journal of Politics* 73(03):692–706.

Gomez, Brad T. and J. Matthew Wilson. 2006. "Rethinking Symbolic Racism: Evidence of Attribution Bias." *Journal of Politics* 68(3):611–625.

Gregorich, Steven E. 2006. "Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework." *Medical care* 44(11 Suppl 3):S78–S94.

Hagquist, Curt and David Andrich. 2004. "Is the Sense of Coherence-Instrument Applicable on Adolescents? A Latent Trait Analysis Using Rasch-Modelling." *Personality and Individual Differences* 36(4):955–968.

Hagquist, Curt and David Andrich. 2017. "Recent Advances in Analysis of Differential Item Functioning in Health Research Using the Rasch Model." *Health and Quality of Life Outcomes* 15(1):181.

Hajnal, Zoltan and Michael U. Rivera. 2014. "Immigration, Latinos, and White Partisan Politics: The New Democratic Defection." *American Journal of Political Science* 58(4):773–789.

Harden, Jeffrey J., Anand E. Sokhey and Hannah Wilson. 2019. "Replications in Context: A Framework for Evaluating New Methods in Quantitative Political Science." *Political Analysis* 27(1):119–125.

Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll and Keith T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Hatemi, Peter K., Charles Crabtree and Kevin B. Smith. 2019. "Ideology Justifies Morality: Political Beliefs Predict Moral Foundations." *American Journal of Political Science* 63(4):788–806.

Hetherington, Marc and Elizabeth Suhay. 2011. "Authoritarianism, Threat, and Americans' Support for the War on Terror." *American Journal of Political Science* 55(3):546–560.

Hetherington, Marc J. and Jason A. Husser. 2012. "How Trust Matters: The Changing Political Relevance of Political Trust." *American Journal of Political Science* 56(2):312–325.

Hetherington, Marc J. and Jonathan D. Weiler. 2009. *Authoritarianism and Polarization in American Politics*. New York, NY: Cambridge University Press.

Hutchings, Vincent L., Hanes Walton and Andrea Benjamin. 2010. "The Impact of Explicit Racial Cues on Gender Differences in Support for Confederate Symbols and Partisanship." *The Journal of Politics* 72(4):1175–1188.

Jacoby, William G. 1991. *Data Theory and Dimensional Analysis*. Newbury Park, Calif.: SAGE Publications, Inc.

Kam, Cindy D. and Donald R. Kinder. 2012. "Ethnocentrism as a Short-Term Force in the 2008 American Presidential Election." *American Journal of Political Science* 56(2):326–340.

Kinder, Donald R. and Cindy D. Kam. 2010. *Us Against Them: Ethnocentric Foundations of American Opinion*. University of Chicago Press.

Kinder, Donald R. and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. Chicago, IL: University of Chicago Press.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon and Ajay Tandon. 2003. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *The American Political Science Review* 97(4):567–583.

Lelkes, Yphtach and Rebecca Weiss. 2015. "Much Ado about Acquiescence: The Relative Validity and Reliability of Construct-Specific and Agree–Disagree Questions." *Research & Politics* 2(3):2053168015604173.

Lizotte, Mary-Kate and Andrew H. Sidman. 2009. "Explaining the Gender Gap in Political Knowledge." *Politics & Gender* 5(02):127–151.

Malhotra, Neil and Jon A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis* 15(3):286–323.

Masri, Yasmine H. El and David Andrich. 2020. "The Trade-Off between Model Fit, Invariance, and Validity: The Case of PISA Science Assessments." *Applied Measurement in Education* 33(2):174–188.

Miller, Joanne M., Kyle L. Saunders and Christina E. Farhart. 2016. "Conspiracy Endorsement as Motivated Reasoning: The Moderating Roles of Political Knowledge and Trust." *American Journal of Political Science* 60(4):824–844.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Berlin: De Gruyter Mouton.

Mondak, Jeffery J. and Mary R. Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge." *The Journal of Politics* 66(02):492–512.

Mondak, Jeffery J. and Matthew V. Hibbing. 2011. Personality and Public Opinion. In *New Directions in Public Opinion*, ed. Adam J. Berinsky. New York: Routledge.

Narayan, Sowmya and Jon A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60(1):58–88.

O'Brien, Kerry, Walter Forrest, Dermot Lynott and Michael Daly. 2013. "Racism, Gun Ownership and Gun Control: Biased Attitudes in US Whites May Influence Policy Decisions." *PLOS ONE* 8(10):e77552.

Pérez, Efrén O. 2011. "The Origins and Implications of Language Effects in Multilingual Surveys: A MIMIC Approach with Application to Latino Political Attitudes." *Political Analysis* 19(4):434–454.

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. 2013. "An Analysis of ANES Items and Their Use in the Construction of Political Knowledge Scales." *Political Analysis* 21(4):407–429.

Rammstedt, Beatrice, Daniel Danner and Michael Bosnjak. 2017. "Acquiescence Response Styles: A Multilevel Model Explaining Individual-Level and Country-Level Differences." *Personality and Individual Differences* 107:190–194.

Rasch, George. 1980. *Probabilistic Models for Some Intelligence and Achievement Tests*. Expanded edition ed. Chicago, IL: MESA Press.

Robison, Joshua, Randy T. Stevenson, James N. Druckman, Simon Jackman, Jonathan N. Katz and Lynn Vavreck. 2018. "An Audit of Political Behavior Research." *SAGE Open* 8(3):2158244018794769.

Ryan, Timothy J. 2017. "No Compromise: Political Consequences of Moralized Attitudes." *American Journal of Political Science* 61(2):409–423.

Scott, Neil W., Peter M. Fayers, Neil K. Aaronson, Andrew Bottomley, Alexander de Graeff, Mogens Groenvold, Chad Gundy, Michael Koller, Morten A. Petersen and Mirjam A. G. Sprangers. 2009. "A Simulation Study Provided Sample Size Guidance for Differential Item Functioning (DIF) Studies Using Short Scales." *Journal of Clinical Epidemiology* 62(3):288–295.

Sides, John, Michael Tesler and Lynn Vavreck. 2019. *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton University Press.

Stegmueller, Daniel. 2011. "Apples and Oranges? The Problem of Equivalence in Comparative Research." *Political Analysis* 19(4):471–487.

Stenner, Karen. 2005. *The Authoritarian Dynamic*. New York, NY: Cambridge University Press.

Tesler, Michael. 2012. "The Spillover of Racialization into Health Care: How President Obama Polarized Public Opinion by Racial Attitudes and Race." *American Journal of Political Science* 56(3):690–704.

The American National Election Studies. 2016. "The ANES 2012 Time Series Study [Dataset]." Version 5/4/2016. Stanford University & the University of Michigan [Producers]. http://www.electionstudies.org.".

The American National Election Studies. 2017. "The ANES 2016 Time Series Study [Dataset]" Version 5/2/2017. University of Michigan & Stanford University [Producers]. http://www.electionstudies.org.".

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.

Wright, Benjamin D. 1999. Fundamental Measurement for Psychology. In *The New Rules of Measurement: What Every Psychologist and Educator Should Know*, ed. Susan E. Embretson and Scott L. Hershberger. Hillsdale, NJ: Lawrence Erlbaum pp. 65–104.

Figure 1: A workflow for establishing measurement equivalence

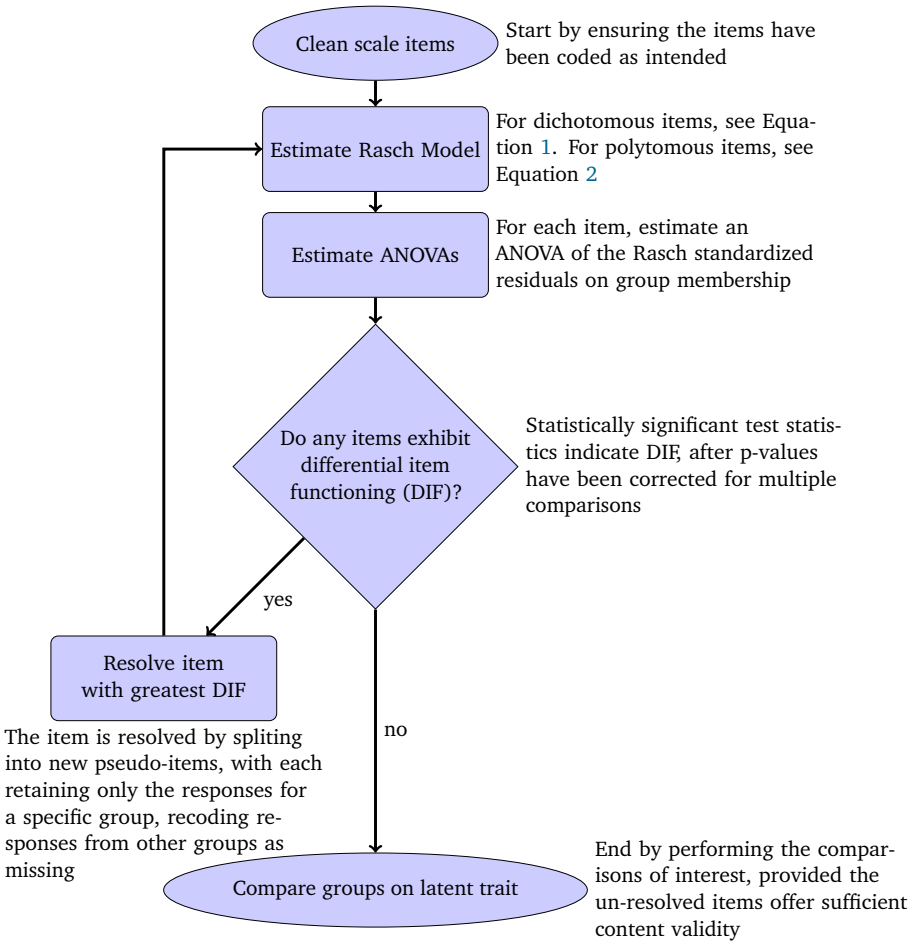
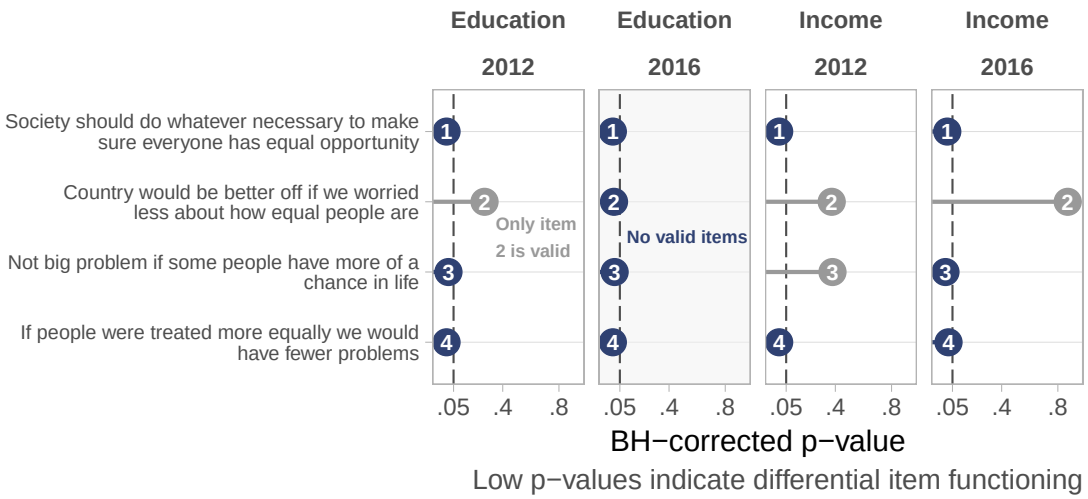
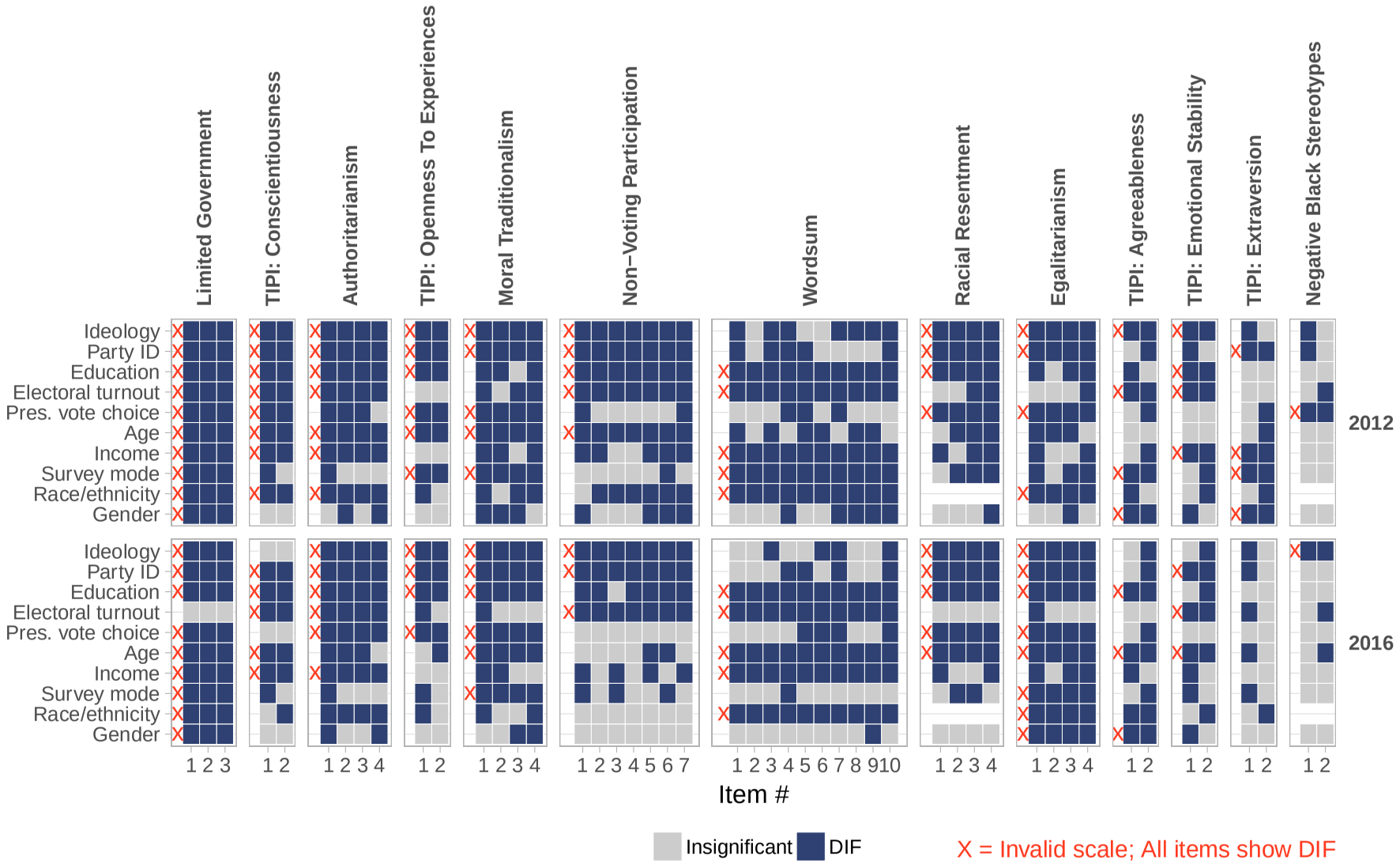


Figure 2: Most items from the Egalitarianism scale exhibit differential item functioning for education and income



Note: The figure displays each Egalitarianism item's p-value from the final ANOVA in which it was included before correction. The ANOVA p-values are corrected for multiple comparisons using the Benjamini and Hochberg (1995) method. For education, the scale shows significant DIF for items 1, 3, and 4 in 2012 and all four items in 2016. For income, the scale shows significant DIF for items 1 and 4 in 2012 and items 1, 3, and 4 in 2016.

Figure 3: DIF tests by scale and grouping variable



Note: The plot shows whether an item exhibited significant DIF in the final ANOVA in which the item was included before correction. For example, the 2012 Authoritarianism scale shows significant DIF by gender for items 2 and 4. The X symbols indicate scale-grouping variable combinations for which the scale is invalid because all items show DIF. For example, the 2012 Authoritarianism-Ideology combination receives an X because it shows significant DIF for all items. The scales measuring Racial Resentment and Negative Black Stereotypes were tested only among white, non-Hispanic respondents. The scales and grouping variables are ordered by the proportion of items showing DIF in 2012.

Figure 4

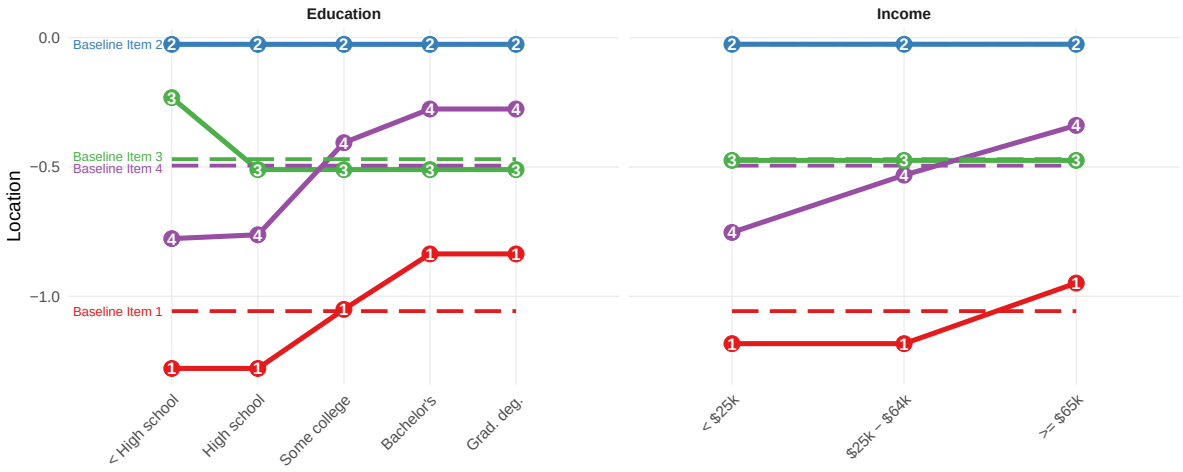


Figure 4A: Egalitarianism item locations vary with education and income

Note: The figure displays how the Egalitarianism item locations vary across education and income levels in the 2012 ANES. An item's location is inversely related to how likely someone is to choose the more egalitarianism response when choosing between two adjacent response categories. The dashed lines represent the locations from the baseline Rasch partial credit model, which assumes no DIF is present. The solid lines represent the locations from the Rasch partial credit model after DIF has been corrected. The corrections are possible because item 2 shows no DIF for education and items 2 and 3 show no DIF for income, providing an anchor to identify the relative locations of the other items for each group.

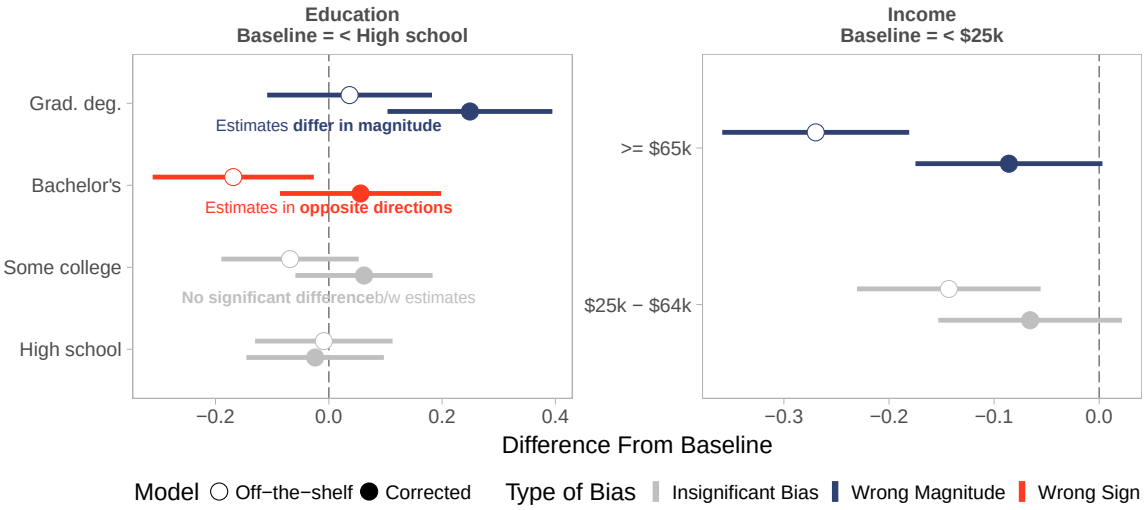
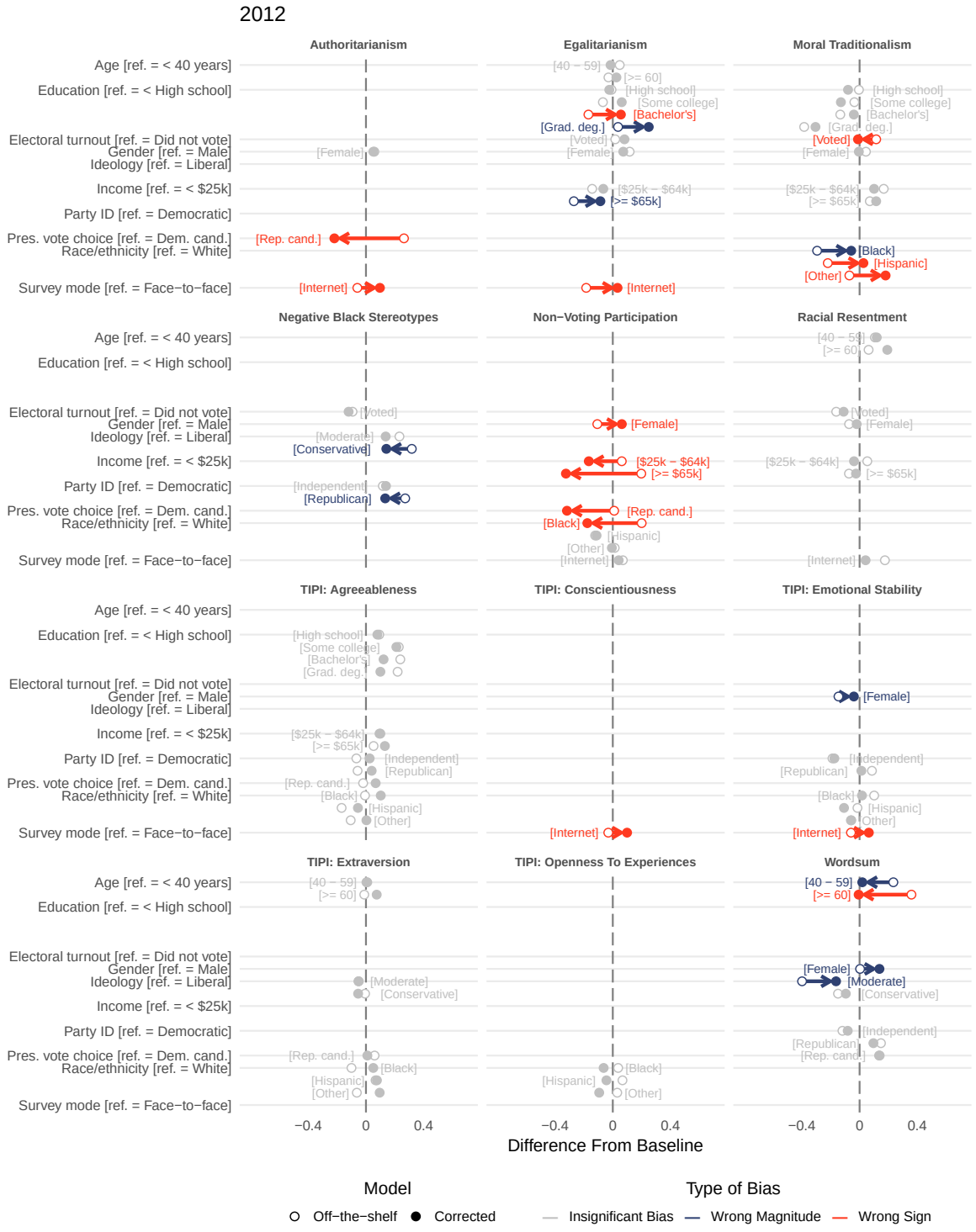
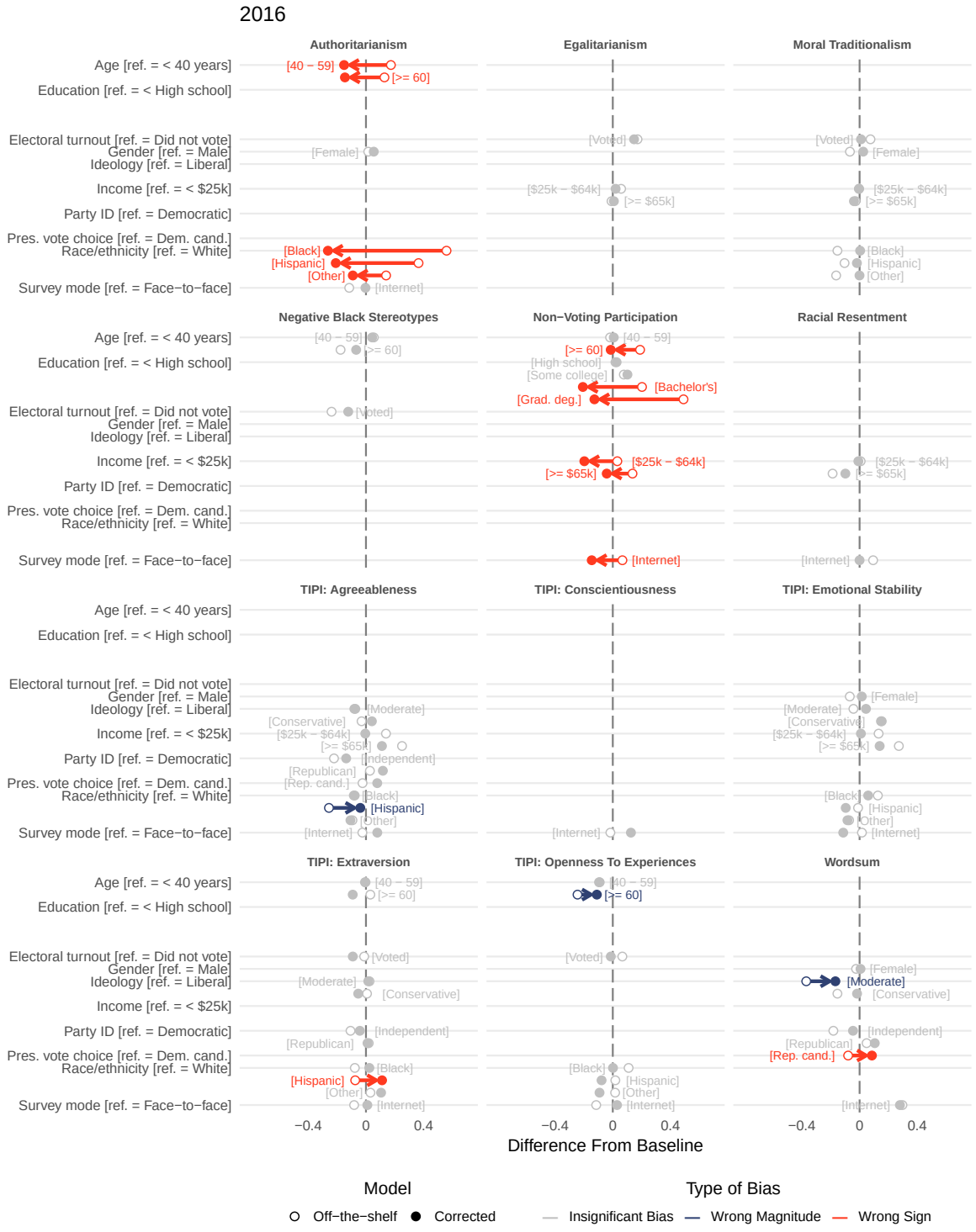


Figure 4B: The 2012 off-the-shelf Egalitarianism scale produces biased estimates of the relationship between egalitarianism and education and the relationship between egalitarianism and income

Note: The open circle displays the off-the-shelf scale's estimated difference between the focal and reference groups. The closed circle displays this difference for the corrected scale. The colored lines indicate a statistically significant difference between the off-the-shelf and corrected estimates. Grey lines indicate that this difference is *not* statistically significant. The figure shows that the off-the-shelf scale produces biased estimates of the difference between the most and least educated individuals (left panel) and the most and least wealthy individuals (right panel).

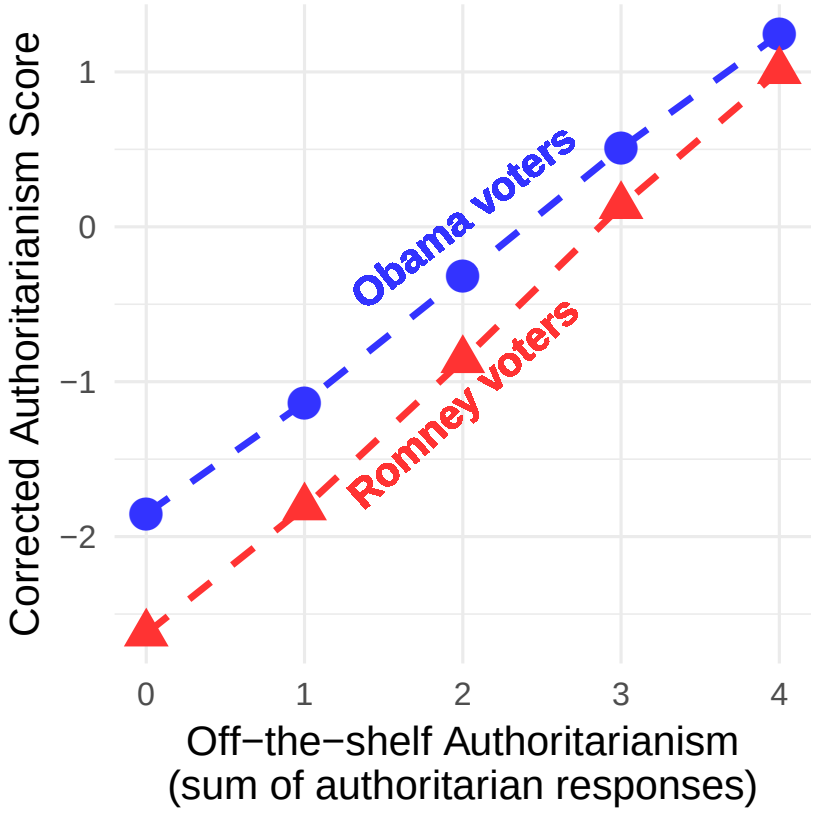
Figure 5: The off-the-shelf scales often suggest different substantive conclusions than do their corrected counterparts.





Note: The open circle displays the off-the-shelf scale's estimated difference between the focal and reference groups. The closed circle displays this difference for the corrected scale. An arrow indicates a statistically significant difference between these estimates. The estimates are not displayed if a valid correction was not possible or if the scale exhibited no DIF.

Figure 6: Obama voters are more authoritarian than Romney voters who receive the same off-the-shelf score.



Note: The figure shows the expected value of the latent trait from the corrected scale given the number of authoritarian responses on the off-the-shelf scale. At every level of off-the-shelf authoritarianism, Obama voters tend to have higher values of the latent trait than Romney voters. The scale lacks measurement equivalence because Romney voters are expected to receive a different score on the off-the-shelf scale than Obama voters with the same level of the latent trait. Since Romney voters receive higher off-the-shelf scores than equally authoritarian Obama voters, the off-the-shelf scale exaggerates Romney voters' value of the latent trait relative to Obama voters.